Human Genome Project

MITRE

Human Genome Project

Study Leader:
S. Koonin

Contributors Include:
S. Block
J. Cornwall
W. Dally
F. Dyson
N. Fortson
G. Joyce
H. J. Kimble
N. Lewis
C. Max
T. Prince
R. Schwitters
P. Weinberger
W. H. Woodin

January 1998

JSR-97-315

Approved for public release; distribution unlimited.

JASON
The MITRE Corporation
1820 Dolley Madison Boulevard
McLean, Virginia 22102-3481
(703) 883-6997

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

Public reporting burden for this collection of information estimated to average 1 hour per response, including the time for review instructions, searc hing existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget. Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
January 4, 1998

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Human Genome Project

5. FUNDING NUMBERS
13-958534-04

6. AUTHOR(S)
S. Koonin, S. Block, J. Cornwall, W. Dally, F. Dyson, N. Fortson, G. Joyce, H. J.Kimble, N. Lewis, C. Max, T. Prince, R. Schwitters, P. Weinberger, W.H. Woodin

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The MITRE Corporation
JASON Program Office
1820 Dolley Madison Blvd
McLean, Virginia 22102

8. PERFORMING ORGANIZATION REPORT NUMBER
JSR-97-315

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Department of Energy
Research and Development
Defense Programs
Washington, DC 20585

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
JSR-97-315

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE
Distribution Statement A

13. ABSTRACT *(Maximum 200 words)*

The study reviews Department of Energy supported aspects of the United States Human Genome Project, the joint National Institutes of Health/Department of Energy program to characterize all human genetic material, to discover the set of human genes, and to render them accessible for further biological study. The study concentrates on issues of technology, quality assurance/control, and informatics relevant to current effort on the genome project and needs beyond it. Recommendations are presented on areas of the genome program that are of particular interest to and supported by the Department of Energy.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

Contents

1 BACKGROUND, CHARGE, AND RECOMMENDATIONS 1

1.1 Overview of the Human Genome Project 1

1.2 Challenges for the Project 2

1.2.1 The complexity of genomic data 2

1.2.2 The state of technology 3

1.2.3 The pace of sequencing 4

1.2.4 The cost of sequencing 7

1.2.5 Project coordination 7

1.3 Study Charge 8

1.4 Recommendations 9

1.4.1 General recommendations 9

1.4.2 Technology recommendations 9

1.4.3 Quality recommendations 11

1.4.4 Informatics recommendations 12

2 TECHNOLOGY 13

2.1 Improvements of Present Genomics Recommendations 13

2.1.1 Electrophoresis improvements and an ABI users group recommendations 14

2.1.2 Algorithms 16

2.2 DOE's Mission for Advanced Sequencing Technology 19

2.2.1 Institutional barriers to advanced technology development 21

2.2.2 Purposes of advanced sequencing technology 23

2.3 Specific Advanced Technologies 24

2.3.1 Single-molecule sequencing 25

2.3.2 Mass-spectrometric sequencing 31

2.3.3 Hybridization 35

3 QUALITY 39

3.1 Quality Requirements 40

3.1.1 The diversity of quality requirements 40

3.1.2 Accuracy required for assembly 42

3.2 Verification Protocols 46

3.2.1 Restriction enzyme verification of sequence accuracy . 47
3.2.2 Hybridization arrays for sequence verification 51
3.2.3 Implementation of verification protocols 55
3.3 Assessing and Improving Present Techniques 55
3.3.1 A systems approach is required 56
3.3.2 Gold standards for measuring sequence accuracy 58
3.3.3 Quality issues pertaining to sequencing templates . . . 60

4 GENOME INFORMATICS 63
4.1 Introduction 63
4.2 Databases 67
4.2.1 User issues 69
4.2.2 Modularity and standards 70
4.2.3 Scaling and storage 72
4.2.4 Archiving raw data 73
4.3 Measures of Success 75
4.4 Sociological Issues 75

1 BACKGROUND, CHARGE, AND RECOMMENDATIONS

1.1 Overview of the Human Genome Project

The US Human Genome Project (the "Project") is a joint DOE/NIH effort that was formally initiated in 1990. Its stated goal is

> *"...to characterize all the human genetic material—the genome—by improving existing human genetic maps, constructing physical maps of entire chromosomes, and ultimately determining the complete sequence...to discover all of the more than 50,000 human genes and render them accessible for further biological study."*

The original 5-year plan was updated and modified in 1993 [F. Collins and D. Galas, "A new five-year plan for the US Human Genome Project," Science **262**, 43-46 (1993)]. The Project's goals to be achieved by the end of FY98 that are relevant for this study are:

- To complete an STS (Sequence Tagged Site) map of the entire genome at 100 kb resolution
- To develop approaches for sequencing Mb regions
- To develop technology for high-throughput sequencing, considering the process as integral from template preparation to data analysis.
- To achieve a large-scale sequencing capacity of 50 Mb/yr and to have completed 80 Mb of human sequence
- To develop methods for identifying and locating genes

- To develop and disseminate software to archive, compare, and interpret genomic data .

Congress has authorized funding through the planned completion of the Project in FY05. The funding in FY97 is $189M for the NIH activity and $78M for the DOE. Thus the total US effort is $267M this year. This amounts to more than half of the worldwide effort, with France, UK, the EU, and Japan being the other major partners.

The DOE program in FY97 included $29M for production sequencing, $15M for the startup of the Joint Genome Institute (a "factory scale" sequencing facility to be operated jointly by LLNL, LANL, and LBNL), $13M for technology development, $11M for informatics, and $3M for applications (construction of DNA libraries, studying gene function, etc.)

1.2 Challenges for the Project

There are a number of challenges that the Project faces if it is to meet its stated goals. We briefly describe several of them in this section as a background to our charge.

1.2.1 The complexity of genomic data

One of the challenges to understanding the genome is the sheer complexity of genomic data. Not all sequence is equivalent. The 3–5% of the genome that is coding consists of triplet codons that specify amino acid sequence. The control regions are binding sites for regulatory proteins that control gene expression. The functions of the introns within a gene and the intergenic regions are largely unknown, even though they comprise the bulk of the genome. There are also special structural elements (centromeres and telomeres) that have characteristic base patterns.

Even given the sequence, the genes are not manifest. And the function and control of a particular gene (When and where is it expressed? What is the function of the protein it encodes?) generally must be determined from the biological context, information beyond the bare sequence itself.

Yet another challenge is that the genomes of any two individuals (except of identical twins) are different (at the 10^{-3} level in the non-coding region; 3–5 times less in the coding regions), and that the homologies between organisms are invariably less than perfect.

Many of these difficulties arise because we don't yet understand the language of the genome. A good metaphor for the state of genetic information is "It's like going to the opera." That is, it's clear something substantial is happening and oftimes it's quite beautiful. Yet we can't really know what's going on because we don't understand the language.

1.2.2 The state of technology

Another hurdle for the Project is the state of technology. The present state of the art is defined by Sanger sequencing, with fragments labeled by fluorescent dyes and separated in length by gel electrophoresis (EP). A basic deficiency of the present technology is its limited read-length capability (the number of contiguous bases that can be read); best current practice can achieve 700–800 bases, with perhaps 1000 bases being the ultimate limit. Since interesting sequence lengths are much longer than this (40 kb for a cosmid clone, 100 kb or more for a gene), the present technology requires that long lengths of DNA be fragmented into overlapping short segments (~1 kb long) that can be sequenced directly. These shorter reads must then be assembled into the final sequence. Much of the current effort at some sequence centers (up to 50%) goes into the assembly and finishing of sequence (closing gaps, untangling compressions, handling repeats, etc.). Longer read lengths would step up the pace and quality of sequencing, although the problem of compressions would still remain.

However, it is important to realize that, beyond the various genome projects, there is little pressure for longer read lengths. The 500–700 base reads allowed by the current technology are well-suited to many scientific needs (pharmaceutical searches, studies of some polymorphisms, studies of some genetic diseases). Thus, the goal of the entire sequence implies technology needs for which there are limited medical or pharmaceutical drivers.

Other drawbacks of the present technology include the time- and labor-intensive nature of gel preparation and running and the comparatively large sample amounts required to sequence. This latter influences the cost of reagents involved, as well as the necessity for extra PCR steps.

1.2.3 The pace of sequencing

One regularly updated "score card" of the Human Genome Project is maintained at http://weber.u.washington.edu/~ roach/human_genome_progress2.htm. This site regularly updates its tallies from the standard human genome databases. As of 5/1/97, there was some 39 Mb of human sequence in contigs of 10 kb or longer; this has been accumulated over the past 20 years. Although 98.7% of the genome thus remains to be sequenced, 15 Mb have been added in the past year. Figure 1.1 shows the progress in the past few years.

The world's large-scale sequencing capacity is estimated to be roughly 100 Mb/yr; although not all of this resource is applied to the human genome. The Joint Genome Institute is projected to have a sequencing capacity of 57 Mb/yr in FY98, growing to 250 Mb/yr in FY01. These capacities are to be compared with the Project's 9/98 goal of 50 Mb/yr.

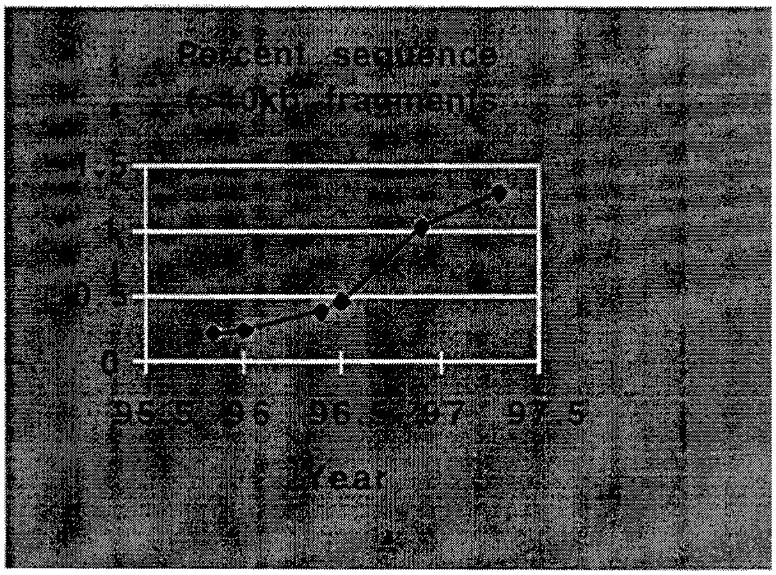

Figure 1.1 Fraction of the human genome in contigs longer than 10 kb that is deposited in publically accessible databases.

It is sobering to contemplate that an average production of 400 Mb/yr is required to complete the sequence "on time" (i.e., by FY05); this corresponds to a daily generation of 50,000 samples and 15 Gbytes of raw data (if the EP traces are archived). Alternatively, an FY98 capacity of 50 Mb/yr must double every 18 months over the next seven years. These figures correspond to an ultimate scale-up of the present capacity by a factor of 30–100. Most observers believe that significant technology advances will be required to meet the FY05 goal.

The length of the known human sequences is also important. The Project's goal is the contiguous sequence of the entire genome. The table below (taken from http://weber.u.washington.edu/~roach/human_genome_progress2.htm) shows the number of known contiguous segments that are equal to or greater than a specified cut-off length. Note that only 1/3 of the known sequence is in lengths of 100 kb or greater, and that the longest human contig is about 1 Mb. It should also be noted that there are many known sequences of several hundred bases or less, for cDNA fragments of this size are generated at a prodigious rate by the public Merck-Washington University collaborative effort and in the pharmaceutical industry. (We heard of one company, Incyte, which produces 8 Mb of raw sequence each day, albeit in small fragments.)

Length cutoff (kb)	Contigs longer than cutoff	Sequence in contigs (Mb)
100	112	16.15
50	191	22.06
40	302	26.82
30	494	33.72
20	579	35.85
10	782	38.66
5	1227	41.72
1	5283	50.50
0.1	very many	—

1.2.4 The cost of sequencing

The cost of sequencing is also a major consideration. If funding continues at the present rate over the next 8 years, the US Project will spend some $2.5B. If all of this funding were devoted to production sequencing, not the case today, a cost of roughly $1 per base would suffice.

Several cost benchmarks are available. The tenth complete microbial genome (*Bacillus subtilis*) has just been announced. It consists of 4000 genes in 4.2 Mb of sequence. This joint European/Japanese project cost something over $2 per base sequenced. Best practice in the Human Genome Project is currently $0.5/base, and the project goal is less than $0.10/base. Specific plans for the Joint Genome Center project an initial (FY97) cost of $0.60 per base, falling to $0.10 per base by FY01. It should be noted that there is difficulty in comparing the costs claimed across laboratories, and across the different funding systems in different nations.

1.2.5 Project coordination

The Human Genome Project presents an unprecedented set of organizational challenges for the biology community. Success will require setting objective and quantitative standards for sequencing costs (capital, labor, and operations) and sequencing output (error rate, continuity, and amount). It will also require coordinating the efforts of many laboratories of varying sizes supported by multiple funding sources in the US and abroad.

A number of diverse scientific fields have successfully adapted to a "Big Science" mode of operation (nuclear and particle physics, space and planetary science, astronomy, and oceanography being among the prominent examples). Such transitions have not been easy on the scientists involved. However, in essentially all cases the need to construct and allocate scarce facilities has been an important organizing factor. No such centripetal force is apparent (or likely) in the genomics community, although the Project would likely benefit from the coordination it would produce.

1.3 Study Charge

Our study was focused on three broad areas:

- **Technology**: Survey the state-of-the-art in sequencing. What are the alternatives beyond gel electrophoresis? What strategies should be used for inserting new technologies into production sequencing? What are the broader uses of sequencing technologies? What are the technology needs beyond those of the Human Genome Project?
- **Quality Assurance and Quality Control**: What are the end-to-end QA/QC issues and needs of the Human Genome Project? What levels of sequence quality are required by various users of genome data? What steps can be taken to ensure these various levels of quality?
- **Informatics**: Survey the current database issues, including data integrity, submission, annotation, and usability? What is the current state of algorithm development for finishing and annotating sequence?

Beyond briefings focused on these specific topics, we also heard a variety of speakers on functional genomics, in order to better get a sense of the needs, standards, and expectations of the consumers of genomic information.

Our recommendations in response to this charge are given in the following section. The balance of this report provides the necessary context and detail, dealing successively with Technology (Section 2), Quality (Section 3), and Informatics (Section 4).

1.4 Recommendations

1.4.1 General recommendations

We begin with two recommendations pertinent to many aspects of the Human Genome Project.

"Know thy system". It is important to have a comprehensive, intimate, and detailed understanding of the sequencing process and the uses of genomic data. Gaining such understanding is quite a different exercise from sequencing itself. Answers to questions such as "What are the pacing factors in production sequencing?"(cloning? gel prep? run time?, lane count?, read length?, ...) or "What is the sequence error budget?" or "What quality of sequence is required?" are essential to optimizing the Project's utility and use of resources.

Couple users/providers of technology, sequence, data. The Human Genome Project involves technology development, production sequencing, and sequence utilization. Greater coupling of these three areas can only improve the Project. Technology development should be coordinated with the needs and problems of production sequencing, while sequence generation and informatics tools must address the needs of data users. Promotion of such coupling is an important role for the funding agencies.

1.4.2 Technology recommendations

Technology development should be emphasized as a DOE strength. Technology development is essential if the Human Genome Project is to meet its cost, schedule, and quality goals. DOE technology development leverages traditional and extensive Department expertise in the physical sciences, engineering, and the life sciences. These are, in many ways, complementary to NIH strengths and interests. If the DOE does not continue to play a leading role in technology development for high-throughput, high-capacity sequencing, it is not clear to us who will.

Continue work to improve present technologies. Although a number of advanced sequencing technologies look promising, none are sufficiently mature to be candidates for the near-term major scale-up needed. Thus, it is important to support research aimed at improving the present Sanger/EP effort. There are clear hardware and software opportunities for improving gel reading capabilities; formation of an ABI user group might accelerate the realization and dissemination of these improvements. There are also software opportunities to improve the crucial assembly and finishing processes, for example by developing a common set of finishing rules, as discussed in Section 2.1.2.

Enhance long-term technology research. The present sequencing technology leaves much to be desired and *must* be supplanted in the long term if the potential for genomic science is to be fully realized. Promising directions at present for advanced technology development include single-molecule sequencing, mass spectrometric methods, hybridization arrays, and micro-fluidic capabilities. The total FY97 funding for advanced technology (i.e., non-EP based) was only $1.7M of the roughly $13M total technology funding in the $78M DOE Human Genome Project; it should be increased by approximately 50%.

Retain technology flexibility in production sequencing facilities. Because sequencing technology should (and is likely to) evolve rapidly (ideally, both evolutionary and revolutionary changes will occur before FY05) it is important to retain the flexibility to insert new technologies into the large-scale sequencing operations now being created (e.g., the Joint Genome Institute). The decisions of when to freeze technology and how much upgrade flexibility to retain are faced in most large scientific projects (e.g., spacecraft or accelerators) and, unfortunately we have no magic prescription for dealing with them. However, the common sense steps of building in modularity and of thoroughly and frequently scanning the technology horizon are well worth remembering.

1.4.3 Quality recommendations

Work to make quality considerations an integral part of the Project. Quality issues must be brought to the fore in the sequencing community, since measures of sequence quality will greatly enhance the utility of the Human Genome Project's "product." Among the top-level steps that should be taken are allocating resources specifically for quality issues and establishing a separate QA/QC research program (perhaps a group at each sequencing center).

Quantify QA/QC issues. Promote research aimed at quantifying (through simulation and other methods) the accuracy required by various end uses of genomic data. Further, since accuracy is a full-systems issue, there is the need for a comprehensive, end-to-end analysis of the error budget and error propagation in the sequencing process, from clone library development through sequencing to databases and analysis software. "You can't discuss it if you can't quantify it."

Develop and implement QA/QC protocols. Develop, distribute, and use "gold standard" DNA as tests of sequencing centers (Section 3.3.2). Support research aimed at developing, validating, and implementing useful verification protocols, along the lines discussed in Section 3.2. Make quality assessments an integral part of all database sequence. A good start would be to require that all database entries include quality scores for each base call. Existing sequencing software tools such as PHRED, PHRAP, and CONSED produce figures of merit for base calls and DNA assembly. While there is room for innovative research aimed at improving the basis for these figures of merit, the existing confidence indicators are nevertheless quite informative and should be made available to users of sequence data.

1.4.4 Informatics recommendations

Listen to the customers. Adhere to a "bottom-up", "customer" approach to informatics efforts supported by DOE. Encourage forums, including close collaborative programs, between the users and providers of DOE-supported informatics tools, with the purposes of determining what tools are needed and of training researchers in the use of new tools and methods. Further, critically evaluate DOE-supported informatics centers with regards to the actual use of their information and services by the community.

Encourage standardization. Encourage the standardization of data formats, software components and nomenclature across the community. Invest in translators if multiple formats exist. Modularize the functions of data archiving, data retrieval, and data manipulation. Distribute the effort for development across several groups. Standardization of data formats allows more than one group to work in each area.

Maintain flexibility. Do *not* demand that "one-size" (in databases) fits all. Make it easy to perform the most common operations and queries, but do not make it impossible for the expert user to execute complicated operations on the data. The community should be supporting several database efforts and promoting standardized interfaces and tools among those efforts.

2 TECHNOLOGY

The technology to sequence the human genome is now in hand. Indeed, this was true when the Project was formulated and initiated in 1990, and there have been significant improvements in the intervening seven years. Nevertheless, as we have noted in Sections 1.2.2-4, there are ample reasons to improve the present technology, particularly if the Project's cost, schedule, and quality goals are to be achieved. Further, improvements in sequencing technology will accelerate genomics research and applications beyond human biology and medicine.

The Project faces the classic dilemma inherent in any large technological project: when to freeze the technology available, to declare "good enough" at the risk of not pursuing the "better." We believe that the likely inadequacy and ease of improvement of the present technology and the future importance and relative inexpense of developing radically different technology all argue for pursuing both tracks simultaneously. Our rationale is presented in the following sections.

2.1 Improvements of Present Genomics Recommendations

In the course of our study, we identified two aspects of the present sequencing technology where improvements that could have a significant impact seemed possible. These are:

- Electrophoresis
- Algorithms for base calling, assembly, and finishing

We consider these in turn.

2.1.1 Electrophoresis improvements and an ABI users group recommendations

The Applied Biosystems Inc. (ABI) automated DNA sequencers are the *de facto* standard for sequencing and will almost certainly carry the brunt of the sequencing load for the Project. These are "closed-box" instruments that utilize proprietary technology owned exclusively by ABI. The company has both the responsibility and the financial incentive to ensure reliable, standardized operation of its instruments, even if this results in sequencing that is less than optimal. On the other hand, the desire of many end users, especially those at major genome sequencing centers, is to push the performance of these instruments to the limit.

This tension raises issues both of technology *per se* and of how new technology can be inserted into ABI machines to the satisfaction of all. We first discuss possible technology improvements, then propose a users group.

It is clear that modifications could be made to the hardware, and especially the software, of the ABI sequencers without sacrificing accuracy of base calling or reliability of operation; one of our briefers spoke convincingly to this issue [C. Tibbets, briefing to JASON, July 1, 1997]. These instruments use the Sanger sequencing method to sample automatically molecules labeled with any of four (ABI-proprietary) fluorescent dyes. The samples undergo gel EP in 36 lanes. The lanes are scanned with an argon laser and bases are "called" by a combination of hardware and software.

Errors can (and do) arise from a number of sources, including lane tracking; differential migration of the four dyes; overlapping emission spectra of the dyes; and variable oligomer separations, due, for example, to secondary sources. There are a number of efforts underway to improve the software packages used for interpreting the (trace) data stream produced by the sequencing instrument. It is important to note that specific improvements might have a dramatic impact on the Project, but be of marginal significance for broad classes of commercial applications. One example is attaining longer read lengths.

Specific areas with clear potential for significant improvement include:

- increasing the lateral scan resolution, thus allowing for more lanes;
- indexing the lateral scan in space (instead of time) for greater trace precision and reproducibility;
- adding a fifth dye for enhanced lane tracking;
- allowing access to the raw (preprocessed) trace data, thus enabling improved base calling algorithms.

ABI has no obligation to respond to users' requests for modifications such as those suggested above, nor are they required to make available detailed specifications that would allow users to make such modifications themselves. As a result, advanced users are taking matters into their own hands through reverse engineering, even if this risks invalidating the manufacturer's warranty or service agreement. For both legal and sociological reasons these aftermarket modifications tend to be made at the level of individual genome centers. This may result in fragmentation of the standards of practice for acquisition of sequence data, complicating the establishment of quality-control measures across the entire genomics community.

It would be desirable to unify the genomics community's efforts to enhance the performance of ABI instruments, without infringing on ABI's right to control its products and to guard its proprietary technology. We recommend that DOE take an active role in setting up an ABI "users group" that would serve as a sounding board for issues pertaining to the operation of existing instruments, the modification of existing instruments for enhanced performance, and the development of next-generation instruments. The group would include members from each of the major genome centers, various private genomics companies that choose to participate, and a sampling of small-scale users who receive federal support for DNA sequencing activities. The group should also include a representative from DOE, NIH, and (if it wishes to participate) ABI itself.

The activities of the users' group should be self-determined, but might include in-person or electronic meetings, generation of reports or recommendations concerning the operation and potential improvement of the ABI instruments, and distribution of information to the scientific community via journal articles or the World Wide Web. DOE should provide principal funding for these activities, although industry members and ABI should pay expenses related to their own participation. It must be understood by all participants that ABI is under no obligation to consider or follow the recommendations of the users' group. We would expect, however, that by finding common ground and speaking with one voice, the users will have substantial impact on the improvement of automated DNA sequencing technology, while maintaining common standards of practice across the genomics field and respecting the proprietary rights to sequencing technology.

2.1.2 Algorithms

Algorithms (and the software packages in which they are embodied) for lane tracking, base calling, assembly, and finishing appear to be in a formative stage. Research into new algorithms, and development and dissemination of software packages containing them, can return significant dividends in terms of both productivity and accuracy.

2.1.2.1 *Base calling*

The base calling problem involves converting a four-channel record of dye fluorescence intensity to a sequence of bases, along with a confidence value for each base. Several factors make this a challenging problem. Spreading of the intensity function along the lane leads to inter-symbol interference. Overlap in the spectral response of the four dyes leads to cross-talk. The spacing between bases may be non-uniform, certain sequences of bases distort the record, and the signal levels are very low toward the end of a read.

All of the problems present in base calling are also present in the demodulation of signals in communication and magnetic recording systems. As a result, there is a rich literature of methods for dealing with these problem. For example, inter-symbol interference can be reduced by employing linear equalization or decision-feedback equalization. Clock-recovery methods can be applied to keep the base calls properly centered. Sequences can be decoded as multi-base symbols to compensate for sequence-dependent distortion. A trellis decoder or a hidden Markov model can be employed to exploit knowledge about expected sequences to compute the most likely sequence to be generated by a particular intensity record. It would be worthwhile to consider implementing new (or improving present) base calling algorithms on the basis of these techniques.

2.1.2.2 *Assembly*

Assembly algorithms stitch together a set of sequences (of perhaps 500 bases each) that are subsequences of a clone (of perhaps 30 kb in length) to generate the (hopefully) complete sequence of the clone. The process is similar to assembling a linear puzzle where the pieces are allowed to overlap arbitrarily. We saw considerable variability in the methods used for assembly. The PHRAP program uses a *greedy* algorithm where the segments with the closest matches are assembled first and the program builds out from this initial start, revising as nececessary. The group at Whitehead, on the other hand, uses an algorithm based on tags to find overlapping segments. All of these algorithms are heuristic and approximate, as a complete search for the optimum map is perceived to require excessive computation.

There are many directions for research on assembly algorithms. To start, better methods for comparing two sequences to determine if they match can be employed. The PHRAP program achieves more accurate assembly by using base-call confidence values in grading matches. This corresponds exactly to the use of soft-decision decoding in a communication system. One can further improve the accuracy of matching by taking into account the sequence-dependent probability of erasures and insertions, computing, for example, the probability of a compression based on the surrounding GC-rich sequence. Similar techniques can be used to handle assembly in the presence of repeats.

Better methods for searching the space of possible assemblies can also be developed. For example, the greedy algorithm employed by PHRAP can get stuck if it makes a wrong choice early in its processing. One should benchmark such algorithms against a complete branch-and-bound search on representative difficult sequences to determine how often such failures occur. If there is a significant advantage to a full search, one can construct special-purpose assembly computers to perform this computation in a reasonable amount of time. For example, one could use an ASIC (Application Specific Integrated Circuit) or a few FPGAs (Field Programmable Gate Arrays) to build an accelerator that plugs into a standard workstation that will compute (in less than a microsecond) matching scores for all shifts of two segments through an algorithm that employs confidence values and sequence-dependent insertions and deletions. Even with a complete search, the use of heuristics is important to guide the search to explore the most likely assemblies first, so that large parts of the search space can be pruned.

2.1.2.3 *Finishing*

The finishing process involves taking an assembled sequence and filling in the gaps through a combination of manual editing and time-consuming directed sequencing. At some sequencing centers, finishing accounts for roughly half of the entire sequencing effort. The software available to assist finishing was of variable sophistication and in some cases consisted of no more than simple sequence editors. While most of the finishing costs are associated with directed sequencing, research into finishing software could help to automate this process.

The first step toward automated finishing is to improve assembly software. Generating a correct assembly without manual intervention would eliminate much of the need for manual editing, leaving only the genuine gaps or compressions to be dealt with using directed sequencing.

The directed sequencing process involves ordering new reads of the clone using primers designed to extend the ends of sections that have already been sequenced. Much of this process can be automated using a rule-based expert system. Such a system is built by having a knowledge engineer observe an expert finisher at work and capture the finisher's thought process in a set of rules. For example,

> *when a contig of a particular length is terminated in a particular way at each end, order a set of primers that match part of the sequence and order new reads taken using these primers and dye-terminator sequencing.*

By combining the approaches taken by several finishers from different centers, the system could, in some cases, outperform a single human finisher. At the very least, a set of a few hundred of these rules would be likely to cover most of the common finishing cases. This would allow the human experts to focus their effort only on the most difficult cases.

2.2 DOE's Mission for Advanced Sequencing Technology

We heard briefings from nine experts describing various technologies that might bring radical improvements to the art of sequencing DNA. These are discussed in some detail below. They are all different, but they have several features in common: they are non-EP, small-scale, and currently absorb a small fraction of the DOE genome project budget (some $1.7M of the $13M DOE technology development budget); unfortunately, they are scheduled to receive even less in the future. These projects are long-range, aimed at developing technologies whose greatest use will be come in the sequel of applications following the initial sequencing of the human genome. They are, to some extent, high-risk, exploring ways to overcome obstacles that could prove to be insuperable. But they also are high-promise, offering a real possibility of new sequencing methods that would be significantly faster and cheaper than gel EP.

How much money should DOE spend on high-risk, high-promise ventures? This is one of the important questions addressed by our study. We recommend a gradual increase of funding for technology development by about 50% (to $20M per year) with a substantial fraction of this money going to projects other than improvements in current gel EP techniques. One should be prepared to increase this level rapidly in case one or more of the new technologies becomes ripe for large-scale operation.

In making this recommendation for increased support for advanced technologies, we are well aware of the need for the DOE to play a significant role in the current stage of the Project. We also know of, and approve of, the technology goals of vastly improving current EP techniques by such means as high-voltage capillaries, ultrathin gels, and use of resonance ionization spectroscopy. It is likely that such improvements in gel EP are essential to completing the Project on time, and we have commented in Section 2.1 on improving gel EP throughput in the near term. However, we believe that in the long run DOE's greatest impact will be in support of the development of advanced technology for various sequencing tasks that go beyond the current goals of faster gel EP.

There are two main reasons for DOE to support these high-risk technologies. First, this is the part of the Project that plays to DOE's strengths: the history and traditions of DOE make it appropriate (indeed, natural) for the Department to explore new sequencing technologies based on the physical sciences. Second, existing gel EP technology is barely adequate for sequencing a single human genome, and new technologies will be required to satisfy the future needs of medicine, biological research, and environmental monitoring. The new ventures supported by DOE are the seed-corn of sequencing efforts, for a crop to be reaped far beyond the Project itself.

2.2.1 Institutional barriers to advanced technology development

Most of the current attention in the Project is currently focused on rapid, low-cost sequencing of a representative human genome, to be finished by FY05. As a result, there has been a tendency to freeze technology at a fairly early level of development, sometimes not much past the proof-of-principle level, in order to cut down lead times. This tendency is exacerbated by the subsequent commercialization of the technology, making it difficult, for the usual property-rights reasons, to incorporate improvements found by those outside the commercial sector. Even this would not be so bad if it were not that the majority of genome researchers are not oriented toward technology development *per se*, but to the biological research that the technology enables. There is a vicious circle in which lack of technology support leads to an insufficient technology knowledge base among the supported researchers, while this lack of knowledge among peer reviewers leads to a reluctance to support technology development.

2.2.1.1 *A parallel in ultrasound technology development*

Three years ago, a JASON study sponsored by DARPA [H. Abarbanel *et al.*, Biomedical Imaging (JASON Report JSR-94-120, August 1995)] looked at the maturity and sophistication of technology both for ultrasound and for MRI. In both cases the study found concrete examples of the institutional barriers discussed in the previous section. Ultrasound was further behind in advanced technology than MRI, and we will comment only on ultrasound here. The problems of ultrasound are well-known to all who work in it. The transmission medium (flesh and bones) is so irregular that images have very poor quality, interpretable only by those devoting their lifetime to it. In-principle improvements were known, especially the construction of two-dimensional ultrasound arrays to replace the universally-used one-dimensional arrays (which severely degrade the resolution in the direction transverse to the array). But this was a difficult technological challenge, requiring sophisticated engineering beyond the reach of much of the ultrasound community, and not representing an obvious profit potential for the commercial suppliers.

The JASON study found that members of the ultrasound research community were largely limited by the pace of commercial technology development, which was conservative and market-oriented, not research-oriented. In some cases there were ultrasound researchers quite capable of making advances in the technology, but frustrated by the lack of NIH funding. The study recommended that DARPA occupy, at least temporarily, the niche of technology development for ultrasound, which existed because agencies like the NIH were not filling it.

In response to this study, DARPA put a considerable amount of money into advancing ultrasound technology, with emphasis on using (two-dimensional) focal-plane array techniques developed by defense contractors for infrared and other electrooptical arrays. While it is too early to foresee the ultimate impact, it appears that this funding will significantly improve ultrasound technology.

2.2.2 Purposes of advanced sequencing technology

The goal of sequencing 3 billion base pairs of a representative human genome requires a limited amount of redundancy (perhaps a factor of 10) to insure complete coverage and improve accuracy. However, further developments in genomics will have to address questions of diversity, rarity, and genomic function, which may make this sequencing effort seem small.

One can imagine the need for continuing (if not increased) sequencing capacity as diversity becomes the issue. Diversity arises from individual variation (RFLPs, VNTRs, and other manifestations of introns, mutations in genes, etc.) and from the desire to compare human genomes with those of other species, or to compare (parts of) one individual's genome with another's. If it is ever to become possible for MDs and laboratory technicians outside biotechnology laboratories to do sequencing routinely, the sequencing process itself will have to become much simpler, and not subject, for example, to fluctuations in the artistry of the experts who nowadays prepare gels. (Not everyone subscribes to such a goal, the alternative being large sequencing centers to which samples are submitted.). The databases that keep track of this diversity will grow correspondingly, as will the search engines needed to mine the databases. It is not out of the question to anticipate computing needs increasing even faster (a pairwise correlation search of a ten times larger database may require up to one hundred times more searching, for example).

The hunt for rare alleles or rarely expressed genes (associated with rare phenotypes or obscure functions) may call for advanced technology for constructing and searching DNA libraries, perhaps massively-parallel machinery built on a considerably smaller unit scale than is now common.

Functional genomics (to oversimplify, the understanding of the roles and interactions of the proteins coded for by DNA) presents difficulties so specific to each individual case study that it is nearly impossible to summarize here, and we will not attempt to do so. But it is clear that many functional genomics activities will require a total sequencing rate substantially beyond that of the Project.

Advanced technologies also have a role to play in quality assurance and quality control. The chemical and physical bases of current sequencing technology result in intrinsic limitations and susceptibility to errors. Alternative sequencing methodologies at least as accurate and efficient as the present one would allow independent verification of sequence accuracy. An example is given in Section 3.2.2.

Some advanced technology development will be (indeed, is being) done by commercial companies, to be sure, and that is to be welcomed, but if ultrasound or even the current state of the Project is a guide for the future, there is a most important role for DOE advocacy and support of advanced technology beyond the goals of initial sequencing of the human genome.

2.3 Specific Advanced Technologies

One cannot, of course, confidently predict the future of advanced technologies in any area. Instead, we comment in the following subsections on three directions that seem particularly promising:

- Single-molecule sequencing (by STM, AFM, flow cytometry, etc.)
- Mass-spectrometric sequencing
- Massively-parallel sequencing (hybridization arrays)

2.3.1 Single-molecule sequencing

For at least thirty years, some molecular biologists have been dreaming that it might be possible to sequence DNA molecules one at a time. To do this, three steps would need to be taken:

- **Step 1**: stretch out a molecule of DNA in a known orientation, with one end attached to a solid surface and the other end free.
- **Step 2**: detach nucleotides one at a time from the free end.
- **Step 3**: count and identify the nucleotides in order as they are released.

Before any of these three steps were mastered, the technique of sequencing DNA by gel EP was invented and the three steps became unnecessary — gel EP became the standard method of sequencing. A significant disadvantage of this method was the requirement for a macroscopic quantity of identical molecules as input. This requirement initially limited its application to viral genomes and other small pieces of DNA that could be obtained in pure form. The invention of cloning technologies and, subsequently, PCR made the preparation of pure macroscopic quantities of identical molecules routine, allowing gel EP to be applied to all kinds of DNA. Thus, the technology was ready for large-scale development when the Project began (indeed, its availability was one of the factors in initiating the Project) and the technology of single-molecule sequencing was left far behind. [Single-molecule spectroscopy and related fields are nevertheless very active areas of research; see, for example, the *Symposium on Single Molecule Spectroscopy: New Systems and Methods*, held last year in Ascona, Switzerland.]

The Human Genome Project has given relatively small support to some single-molecule sequencing efforts. We heard about only two serious programs to develop single-molecule sequencing. One, at LANL, was described to us in a briefing by Richard Keller. The other, a proprietary program at seQ Ltd. in Princeton, was mentioned, but not described in detail. Neither program is now supported by the Project. Details of the LANL program have been published [P. M. Goodwin, W. P. Ambrose, and R. A. Keller, "Single-molecule Detection in Liquids by Laser-Induced Fluorescence", Accounts of Chemical Research, **29**, 607-613 (1996); R. A. Keller *et al.*, "Single-Molecule Fluorescence Analysis in Solutionî", Applied Spectroscopy, **50**, 12A-32A (1996)]

Why should anybody be interested in single-molecule sequencing? There are two main reasons. First, each of the three steps required for single-molecule sequencing has recently been demonstrated to be feasible. Second, single-molecule sequencing, if all goes well, might turn out to be enormously faster and cheaper than EP. The following paragraphs explain the factual basis for these two statements.

The first step in single-molecule sequencing is the attachment of one end of a molecule to a solid surface and the stretching out of the rest of the molecule in a controlled manner. This has been done by the LANL team, using flow cytometry, a standard technique of microbiology. A single molecule of single-stranded DNA is attached by the covalent bonds of the biotin-avidin protein system to a plastic microsphere. The microsphere is held in an optical trap in a cylindrical fluid flow, which pulls the molecule straight along the cylinder's axis. The second step is the detachment of nucleotides in sequence from the end of the molecule. This has also been demonstrated by the LANL team, using standard microbiological techniques. Exonucleases are dissolved in the flowing fluid. A single exonuclease molecule attaches itself to the free end of the DNA and detaches nucleotides, one at a time, at a rapid rate (many per second).

The third step, the identification of bases in the detached nucleotides, is the most difficult. It might be done in at least three different ways. The LANL team identifies the bases by passing the flowing fluid through a laser-beam. As each base passes though the beam, the molecule fluoresces at a wavelength that is different for each of the four bases. Because the passage through the beam is rapid, the fluorescence must be intense if it is to be detected reliably. To intensify the fluorescence, the DNA molecule is initially prepared for sequencing by attaching a fluorescent dye residue to each base, with four species of dye marking the four species of base. The four types of base can then be identified unambiguously during roughly one millisecond that each nucleotide spends in the laser beam. Unfortunately, the LANL team has not succeeded in eliminating spurious detections arising from unwanted dye molecules in the fluid. They expect to be able to reduce the background of spurious events to a level low enough to allow accurate sequencing, but this remains to be demonstrated; it will require faster-acting exonucleasus than those now used.

The seQ Ltd. team accomplishes the first two steps in the same way as the LANL team, but addresses the third step differently. The bases are not modified by addition of dye residues. Instead, the unmodified nucleotides are detected by fluorescence in an ultraviolet laser-beam. Since the fluorescence of the unmodified bases is relatively weak, they must be exposed to the laser for a longer time. This is achieved by depositing each nucleotide, immediately after it is detached from the DNA, onto a moving solid surface. The surface is then scanned by ultraviolet lasers at a more leisurely pace, so that each nucleotide is exposed to the lasers long enough to be identified unambiguously. The details of this technique are proprietary, and we were not told how well it is actually working.

A third possible way to do the third step in single-molecule sequencing is to use mass spectrometry. The state of the art of mass spectrometry is discussed in Section 2.3.2. Mass-spectrometric identification of the detached nucleotides would require their transfer from the liquid phase into a vacuum. This might be done by ejecting the flowing liquid into a spray of small droplets, letting the droplets evaporate on a solid surface, and then moving the solid surface into a vacuum. Molecules sticking to the surface could then be detached and ionized by MALDI. Once ionized, they could be detected and identified in a mass-spectrograph, since the four species of nucleotide have different masses. (As noted in the next subsection, it is considerably more difficult to differentiate the four base pairs by mass than to distinguish their presence or absence, as in Sanger sequencing.) However, none of the mass-spectrograph projects that we heard about has addressed the problems of single-molecule sequencing.

To summarize the present situation, although each of the steps of single-molecule sequencing has been shown to be feasible, but no group has yet succeeded in putting all three together into a working system. The programs at LANL and seQ Ltd. are on a modest scale. Dr. Keller informs us that he is exploring the possibility of collaboration with a larger German-Swedish consortium headed by Manfred Eigen and Rudolf Rigler. The latter have published a plan for single-molecule sequencing essentially identical to the LANL program [M. Eigen and R. Rigler, Proc. Nat. Acad. Sci. (USA) **91**, 5740 (1994)], although LANL is ahead of the consortium in the implementation of their plan. If the collaboration goes ahead, using the skills of LANL and supportedwill be leveraged by the larger resources of the consortium., there is a good chance that the plan can be developed into a practical system.

We turn now from the present situation to the future promise of single-molecule sequencing. The promise is that it might become radically faster and cheaper than gel electrophoresis. The claim that single-molecule sequencing might be extremely cheap stands or falls with the claim that it might be extremely fast. Sequencing by any method is likely to be a labor-intensive operation, with costs roughly proportional to the number of person-years devoted to it. The costs of machines and materials are likely to be comparable with the costs of wages and salaries. When we are concerned with large-scale operations, the number of bases sequenced per dollar will be roughly proportional to the number of bases sequenced per hour. The main reason why gel electrophoresis is expensive is that it is slow. If single-molecule sequencing can be a hundred times faster than gel electrophoresis, then it is also likely to be a hundred times cheaper.

The claim that single-molecule sequencing might be a hundred times faster than gel electrophoresis rests on a comparison of the inherent speeds of the two processes. The process of gel electrophoresis requires about eight hours to separate molecules with resolution sufficient to sequence 500 bases per lane. The inherent speed of gel electrophoresis is thus less than one base per minute per lane. In contrast, the elementary steps in single-base sequencing might have rates of the order of a hundred bases per second. The digestion of nucleotides in sequence from the end of a DNA molecule by exonuclease enzymes has been observed to occur at rates exceeding one hundred bases per second. And the discrimination of bases in ionized molecules detected by a mass-spectrometer can certainly be done at rates of hundreds of molecules per second. These facts are the basis for hoping that the whole process of single-molecule sequencing might be done at a rate of a hundred bases per second. That would imply that an entire human genome could in principle be sequenced by a single machine operating for a year.

Needless to say, this possibility is very far from being demonstrated. The three steps of single-molecule sequencing have not yet been integrated into a working process. And the rate of sequencing in a large-scale operation is limited by many factors beyond the rates of the elementary process involved. With either single-molecule or gel electrophoresis separation, the production of sequence will be slowed by the complicated manipulations required to prepare the molecules for sequencing and to assemble the sequences afterwards. Until single-molecule sequencing is developed into a complete system, no realistic estimate of its speed and cost can be made. The most that can be claimed is that single-molecule sequencing offers a possibility of radically increasing the speed and radically reducing the cost.

Two other potential advantages of single-base sequencing are longer reading-lengths and superior accuracy. The reading-length in gel EP is limited to about a thousand bases (roughly half of this in conventional practice). The LANL group has demonstrated attachment and suspension of single DNA molecules with many thousand bases. It is likely that DNA molecules with tens of thousands of bases could be handled, so that a single-molecule sequence could have a read length of tens of thousands of bases. As the short read length of gel EP makes final assembly and finishing an elaborate and costly process, these longer reads could greatly simply the process of assembly.

One of the major obstacles to accurate sequencing is the prevalence in the genome of repeated sequences of many kinds. Repeated sequences are a frequent cause of ambiguities and errors in the assembly process. Since the single-molecule system will have longer read lengths, it will be less vulnerable to effects of repetition. Repeated sequences will usually be displayed, without ambiguity, within the compass of a single consecutive read. As a result, it is possible that single-base sequencing may be not only faster, but also more accurate than gel EP.

In addition to the LANL and seQ Ltd. programs and the mass-spectroscopy programs described in the following subsection, there are some efforts directed towards single-molecule sequencing by non-destructive methods using microscopes. The idea of these efforts is to discriminate bases by scanning a DNA molecule with an Atomic Force Microscope or a Scanning Tunneling Microscope. These efforts are much further from practicality than the LANL and seQ Ltd. programs; we have not examined them in detail. Since the art of microscopy is advancing rapidly, it is possible that some new invention will make it possible to visualize individual bases in DNA with enough resolution to tell them apart. However, without a new invention, it appears that the existing microscope technology cannot do the job.

In conclusion, this study's recommendation's that DOE give modest support to single-molecule sequencing efforts. While we have only reviewed two small efforts, it appears to us that, with modest support, there is a chance that single-molecule sequencing will be developed into a practical system within a few years. There is a smaller probability that it will prove to be superior to gel EP by a wide margin. Of course, funding decisions for individual programs, including those we have reviewed, must be made through the usual mechanisms, including rigorous peer review of prior accomplishments and future potential.

One can look at the support of single-molecule sequencing from two points of view. On the one hand, it is a gamble that DOE can afford to take, offering an opportunity to win a large pay-off by betting a small fraction of the genome budget. On the other hand, it is a premium that DOE can afford to pay for insurance against the possibility that the electrophoresis-based sequencing program might fail to reach its schedule, budget, and accuracy goals. From both points of view, modest support of single-molecule sequencing appears to be a prudent investment.

2.3.2 Mass-spectrometric sequencing

In the simplest terms, mass spectrometry (MS) in DNA sequencing replaces the gel EP step in Sanger sequencing. Instead of measuring the lengths of various dideoxy-terminated fragments by observing their rate of diffusion in a gel, one measures their mass with one of several possible MS techniques, including time-of-flight (TOF) and Fourier-transform ion cyclotron resonance (FTICR) spectroscopy. Presently, MS techniques are usable on fragments of about the same length as those used in gel EP (that is, several hundred bases), although this is not a fundamental limitation. The real advantage of MS sequencing is speed, since reading the output of the MS instrument is virtually instantaneous, compared to eight hours or so needed for the gel lanes to evolve to readable length. Many other techniques can be used, in principle, for sequencing with MS, and we will not go into all of them here. Some of these require a mass resolution capable of distinguishing all of the four base pairs by mass; this is a difficult job, since A and T differ by only 9 Da. (Sanger sequencing needs only to resolve one whole base pair, or about 300 Da.)

In early investigations into MS DNA sequencing, the methods for preparing and ionizing DNA (or protein) fragments were fast-atom bombardment or plasma ionization. (There are recent review articles on DNA MS, including references to the work described below [K. K. Murray, J. Mass Spect. **31**, 1203 (1996); P. A. Limbach, Mass Spectrometry Reviews **15**, 297 (1996)]; the discussion here is based on these articles and on remarks from several experts.) But spectroscopy was limited to oligonucleotides of ten or fewer bases.

One significant step forward is the use of MALDI (Matrix-Assisted Laser Desorption/Ionization) to prepare ionic fragments of DNA for MS. The general idea is to embed the DNA in a matrix, which can be as simple as water ice, and to irradiate the complex with a laser of carefully-chosen frequency. This can both vaporize the complex and ionize the DNA, possibly by first ionizing the matrix followed by charge transfer to the DNA. There is a great deal of art in applications of MALDI, which is considerably more difficult to use with DNA than with proteins and peptides. For example, problems arise with unwanted fragmentation of the (already-fragmented) DNA during the MALDI process. Moreover, this MALDI fragmentation process is different for different bases. It is now possible to generate DNA fragments up to 500 bases long with MALDI, with resolution at about the 10 base level (compared to the needed resolution of 1 base). Typically MALDI DNA fragments have one unit of charge for every several hundred base pairs.

Another promising method for ionization is electrospray ionization (ESI). Here the charge produced is much higher (but can be varied by changing the chemistry of the solution containing the DNA). For example, experiments using T4 phage DNA fragments up to 10^8 Da have shown charges up to 3×10^4. It is then necessary to determine both the mass per unit charge (as in conventional TOF MS) *and* the charge, in order to determine the mass. One potentially-important method introduces the accelerated ions into an open metal tube, where they induce an image charge that is measured; the charge-to-mass ratio is then measured by TOF.

MALDI-based methods are generally best for Sanger sequencing, but improvements are needed in the mass resolution and sensitivity (equivalently, DNA ion yield). ESI techniques lead to both higher mass resolution and higher mass accuracy, but because a great many charge states are created, it is not well-suited to analysis of a mixture of a large number of fragments (as is required in Sanger sequencing).

Looking toward the future, there are two ideas in MS that might someday reach fruition.

> **Arrays and multiplex MS sequencing.** Several briefers discussed ideas for using large arrays of DNA fragments with MS. One scheme [Charles Cantor, briefing to JASON, July 3, 1997] involves using arrays with various laydowns of DNA fragments, for subsequent MALDI-MS, with the fragments on the MALDI array designed to have properties desirable for MS. Another [George Church, briefing to JASON, July 2, 1997] points out that multiplexing with arrays is feasible for MS sequencing at rates of possibly 10^3 b/sec. One uses large ($\sim$ 65000) arrays with electrophore-tagged primers on the DNA fragments, with each primer having an electrophore of unique mass attached. DNA primed with these primers is grown with dideoxy terminators, just as in Sanger sequencing. The four varieties are electrophoretically separated, then collected as droplets on an array. Finally, MALDI-TOF is used to remove the electrophores, ionize them, and identify them by MS. Each of the 400 different varieties of DNA is thus identified, yielding a multiplex factor which is the number of different electrophores (400 in this case). (Electrophore tagging of

primers has been suggested as a means of increasing the ion yield from MALDI [P. F. Britt, G. B. Hurst, and M. V. Buchanan, abstract, Human Genome Program Contractor-Grantee Workshop, November ,1994].)

Single-molecule detection. It is not obvious that MS-DNA sequencing requires single-molecule detection, but it in any case can be cited as the ultimate in MS sensitivity. It has already been shown [R. D. Smith *et al.*, Nature **369**, 137 (1994)] that a single ESI-DNA ion (up to 25 kb long) can be isolated for many hours in an FTICR mass spectrometer cell, making it available for measurements during this time. In another direction, detecting a single DNA molecule after acceleration should be possible, thus increasing the sensitivity of MS methods. Methods used for detection might involve bolometric arrays of detectors similar to those used for searches for cosmic dark matter. Such bolometric arrays are made on a pitch of ~ 25 μm for use as sensitive IR focal-plane arrays. An ESI-ionized 30 kDa DNA fragment of charge 100 in a 30 keV potential drop will deposit some 3 MeV in a pixel, the same as 3×10^6 optical photons. The 25 μm spatial resolution can be used for resolving the mass and charge of the ion. It is intriguing to note that a single charged DNA fragment is something like the hypothesized magnetic monopoles of particles physics; both have masses of tens of kDa and large charges (of course, magnetic charge for the monopole). Considerable effort has gone into methods for detection of single monopoles (none of which has been found). [Subsequent to completing this study, we learned of very recent and promising work by Benner et al. at LBNL using superconducting tunnel junctions for single-molecule detection.]

2.3.3 Hybridization

A new technology that has progressed considerably beyond the stage of laboratory research is the construction of large, high density arrays of oligonucleotides arranged in a two-dimensional lattice. ["DNA Sequencing: Massively Parallel Genomics," S. P. A. Fodor, Science **277**, 393 (1997)] In one scheme (termed *Format 1*), DNA fragments (e.g., short clones from DNA libraries) are immobilized at distinct sites on nylon membranes to form arrays of 10^4–10^5 sites with spot-to-spot spacing of roughly 1 mm. ["DNA Sequence Recognition by Hybridization to Short Oligomers: Experimental Verification of the Method on the E. coli Genome," A. Milosavljevic *et al.*, Genomics **37**, 77 (1996)] In a second scheme (termed *Format 2*), techniques of modern photolithography from the semiconductor industry or inkjet technology have been adapted to generate arrays with 400,000 total sites [Fodor, *op cit.*] and densities as high as 10^6 sites/cm^2 ["DNA Sequencing on a Chip," G. Wallraff *et al.*, Chemtech, (February, 1997) 22], although the commercial state of the art appears to be perhaps 10 times smaller. For Format 2 arrays, distinct oligomers (usually termed the *probes*) are lithographically generated *in situ* at each site in the array, with the set of such oligomers designed as part of an overall objective for the array.

In generic terms, operation of the arrays proceeds by interacting the probes with unknown *target* oligonucleotides, with hybridization binding complementary segments of target and probe. For Format 2 arrays, information about binding of target and probe via hybridization at specific sites across an array is obtained via laser excited fluorescence from intercalating dyes which had previously been incorporated into either probe or target, while for Format 1 arrays, readout can be by either phosphor imaging of radioactivity or by fluorescence. Interrogation of the array via changes in conductivity is a promising possibility with potential for both high specificity and integration of the readout hardware onto the array itself [T. Meade, private communication].

Typical probe oligomers are of length 7–20 base pairs, with single base-pair mismatches between target and probe having been detected with good fidelity. ["Mapping Genomic Library Clones Using Oligonucleotide Arrays," R. J. Sapolsky and R. J. Lipshutz, Genomics **33**, 445 (1996); "Accessing Genetic Information with High-Density DNA Arrays," M. Chee *et al.*, Science *274*, 610 (1996)]. For lithographically generated arrays, an important point is that all possible oligomers of length L (of which there are 4^L) can be generated in of order $4L$ processing steps, so that large search spaces (the number of probes) can be created efficiently.

Such large-scale hybridization arrays (with commercial names such *SuperChips* [Hyseq Inc., 670 Almanor Ave., Sunnyvale, CA 94086.] or *GeneChips* [Affymetric, http://www.affymetric.com/research.html] bring a powerful capability for parallel processing to genomic assaying. The list of their demonstrated applications is already impressive and rapidly growing, and includes gene expression studies and DNA sequence determination. While hybridization arrays are in principle capable of *de novo* sequencing ["DNA Sequence Determination by Hybridization: A Strategy for Efficient Large-Scale Sequencing," R. Drmanac *et al.*, Science **260**, 1649(1993)], the combinatorics make this a formidable challenge for long segments of DNA, since an unknown string of length N base pairs is one of p=4^N possibilities (e.g., for $N = 10^3, p \sim 10^{600}$).

Some sense of the probe resource requirements for *de novo* sequencing can be understood by the following "reverse" strategy applied to an array of Format 2 type. Consider an array containing oligomers of total length J with nondegenerate cores of length L that is exposed to an unknown fragment of length N. *A posteriori* one must be left with a sufficient number of probes that have matched the target so that a tiling pattern of probes can be assembled to span the entire target. As a lower bound on the number of required probes, imagine butting a set of N/L probes representing the nondegenerate cores end to end to cover the target, with $p = N/4^L << 1$ so that the conditional probability for two probes to match identical but disjoint regions of the target is small. For $(L, N) = (7, 10^3), p \sim 0.06$, while for $(L, N) = (10, 10^4), p \sim 0.01$. Since each probe has as its nondegenerate segment an arbitrary combination of base pairs, 4^L distinct oligomers are required in the original array, which for $L = 7$ is 2×10^4 elements (well within the realm of current capabilities), while $L = 10$ requires about 10^6 elements (an array with 400,000 sites is the largest of which we are aware).

Unfortunately, this simple strategy does not allow one to deduce the ordering of the matching oligomer segments, of which there are approximately $(N/L)!$ permutations. Hence, imagine augmenting the above strategy so that the matching probes are arranged one after the other with the nondegenerate regions overlapping but offset by k base pairs. That is, adjacent probes are identical to each other and to the target in their overlapping regions, but differ by k base pairs in the nondegenerate regions at each end to provide sufficient redundancy to determine the ordering of the segments with high confidence. The number of probe segments needed to tile the target is then 1+(N-L)/k. With the assumption of only pair-wise probe overlaps (i.e., $k > L/2$), the requirement for uniqueness in sorting then becomes $r = 4^{(L-k)}/[1 + (N - L)/k] >> 1$, which cannot be satisfied for $(L, N) = (7, 10^3)$, while for $(L, N) = (10, 10^3)$, r is at most 5. On the other hand, for sequencing applications with $N = 10^4, L$ must be increased ($L = 14$ gives $r \sim 10$ for $k = 7$), with a concomitant explosion beyond current capabilities in the number of array elements required ($4^{14} = 3 \times 10^8$).

Note that these simple limits assume that target-probe hybridization and identification at each site are perfect and that N is a "typical" random sequence without perverse patterns such as multiple repeats (which would present a significant problem). Certainly in practice a number of processes are encountered that complicate the interpretation of the hybridization patterns presented by arrays (e.g., related to complexity of the thermodynamics of hybridization, of patterns from multiple mismatches, etc.) and that are currently being addressed in the research literature, with promising demonstrations of fidelity. Clearly in any real application somewhat larger arrays than those based upon simple combinatorics will be needed for *de novo* sequencing to maintain accuracy and robustness in the face of errors, with an optimum array size lying somewhere between the limits discussed above.

While there are undoubtedly many "niche" applications for high density hybridization arrays to *de novo* sequencing (e.g., increasing the read length from 500–700 bases to beyond 1 kb would be important in the assembly process), such arrays seem to be better suited to comparative studies that explore differences between probe and target. Indeed, for Format 1 arrays, previously non-sequenced biological materials can be employed. It is clear that hybridization arrays will profoundly impact comparative genetic assays such as in studies of sequence polymorphism [M. Chee *et al.*, op cit.] and of gene identification and expression, as well as for understanding the relationship between genotype and phenotype. Beyond the research environment, one can imagine biochemical micro-laboratories for clinical applications [G. Wallraff *et al.*, op cit.] with hybridization arrays as essential elements for (differential) sequence analysis.

3 QUALITY

A project with the stated goal of sequencing the entire human genome must make data accuracy and data quality integral to its execution. It is clear that much of the genome will later be re-sequenced piece-by-piece. But a high-quality database can reduce the need for such resequencing, provide useful and dense markers across the genome, and enable large-scale statistical studies. A quantitative understanding of data quality across the whole genome sequence is thus almost as important as the sequence itself.

Technology for large-scale DNA sequencing is relatively new. While current sequencing tools and protocols are adequate at the lab-bench level, they are not yet entirely robust. For generic DNA sequence, the mainstream techniques are straightforward and can be carried out with low error rates. However problems and errors occur more frequently when sequencing particular portions of the genome or particular sequence patterns, and resolving them requires expert intervention. Phenomena such as deletions, unremoved vectors, duplicate reads, and chimeras are often the consequence of biological processes, and as such are difficult or impossible to eliminate entirely. Base-call accuracy tends to degrade toward the end of long sequence reads. Assembly of complete genomic sequences remains a challenge, and gaps are sometimes difficult to fill. In this situation, quality assurance and quality control (QA/QC) are essential. In particular it is crucial to understand quantitatively the accuracy of information going into the genome data base. The present section of this report discusses the coupled issues of quality assurance, quality control, and information about data quality, as they impact the Project, as well as other national and international sequencing efforts.

The following three steps provide a useful framework for analyzing and addressing QA/QC issues for the Project (indeed, for any large-scale sequencing effort):

1. Quantify the quality requirements of present and future uses of genomic information

2. Develop assays that can accurately and efficiently measure sequence quality

3. Take steps to ensure that present and evolving sequencing methods and data meet the prescribed level of quality.

The following subsections consider each of these issues in turn. We then follow with some summary recommendations on QA and QC. Following the conclusion of our study, we became aware of a report of an NHGRI Workshop on DNA Sequence Validation held in April, 1996 [http://www.nhgri.nih.gov/HGP/Reports/dna_sequence_workshop.html] that independently examined some of the same issues and came to some similar conclusions.

3.1 Quality Requirements

Our briefers reflected a wide range of opinions on the magnitude of the required error rates for sequence data. This has clearly been a controversial issue and, at times, it has been used as a surrogate for other inter-Center disputes. We believe that the debate on error rates should focus on what level of accuracy is needed for each specific scientific objective or end-use to which the genome data will be put. The necessity of "finishing" the sequence without gaps should be subject to the same considerations. In the present section, we stress the need for developing quantitative accuracy requirements.

3.1.1 The diversity of quality requirements

Genomic data will be (indeed, are being) put to a variety of uses and it is evident that the quality of sequence required varies widely among the possible applications. If we quantify accuracy requirements by the single-base error, $\mathcal{E}$, then we can give some representative estimates:

Application	Error requirement
Assemble long contigs	$\mathcal{E} \sim 10^{-1}$
Identify a 20-mer sequence	$\mathcal{E} \sim 10^{-1}$
Gene finding	$\mathcal{E} \sim 10^{-2}$
Construct a pair of 20-mer STS primers	$\mathcal{E} = 2.5 \times 10^{-4}$ (99% confidence) $\mathcal{E} = 2.5 \times 10^{-3}$ (90% confidence)
Polymorphism	$\mathcal{E} \sim 2.5 \times 10^{-5}$ (coding regions) $\mathcal{E} \sim 10^{-4}$ (non-coding regions)
Studies of genomic evolution, statistics	???

Although these are only rough order-of-magnitude estimates; we justify each as follows.

- The surprisingly low accuracy we estimate to be required to assemble long contigs and to identify the presence of a precisely known 20-mer in a sequence is discussed in the following subsection for the ideal case of no repeats

- Our estimate for the gene finding requirement is based on the observation that pharmaceutical companies engaged in this activity seem satisfied with short sequences (400 bases) at this level of accuracy.

- The required accuracy to construct a pair of 20-mer STS primers is based on straightforward probabilistic calculations.

- The polymorphism entry simply repeats the common statement that accuracy 10 times better than the observed polymorphism rate is sufficient.

- The requirements for evolutionary or statistical studies of the genome have not been quantified

More precise estimates for each of these uses (and others) can surely be generated by researchers expert in each of the various applications. Beyond qualitative judgment, one useful technique would be to run each of the applications with pseudodata in which a test sequence is corrupted by artificially generated errors. Variation of the efficacy of each application with the error level would determine its error requirement and robustness. Such exercises, carried out in software, cost little, yet would go a long way toward setting justifiable quality goals. We recommend that the DOE encourage the genomics community to organize such exercises.

With this kind of data in hand, one could establish global quality requirements for the final sequence (perhaps different for coding and non-coding regions). It is likely that arbitrarily high accuracy could be achieved by expending enough effort: multiple sequencing with alternative technologies could guarantee high accuracy, albeit at unacceptable cost. In the real world, accuracy requirements must be balanced between what the users need, the cost, and the capability of the sequencing technology to deliver a given level of accuracy. Establishing this balance requires an open dialog among the sequence producers, sequence users, and the funding agencies, informed by quantitative analyses.

3.1.2 Accuracy required for assembly

A probabilistic analysis of the assembly problem shows that (in an ideal case) assembly requires relatively little accuracy from the raw sequence data. These data are the sequences of base calls derived from the individual reads. An accuracy as low as 0.9 (per base call) is sufficient to ensure reliable assembly. A high degree of coverage is required, however, to have any chance of assembling the entire clone without gaps. The following analysis justifies these statements, albeit in the absence of repeats, which are likely to complicate the situation considerably.

We first consider the problem of assembling k fragments of length L

with left endpoints uniformly distributed over a clone of length M. Requiring overlaps above a given threshold does not really complicate the gap problem. The point is that a tiling of the sequence of length M with fragments of length L overlapping with subsegments of length at least x is ensured by a tiling with no gaps with fragments of length $L - x$.

We can compute an approximate lower bound for the probability of success as follows. The probability that for a given region of length L^*, some fragment has its left endpoint somewhere in the given region is

$$1 - (1 - L^*/M)^k$$

where k is the number of fragments considered.

We now suppose that the clone length is 30,000 and that the fragments have length 1300. The probability that with 450 fragments there exists a sequence of 150 distinct fragments starting at the left end of the clone such that each successive fragment starts in the left-justified 1200-length subfragment of the previous fragment (thereby ensuring overlaps of 100) is at least

$$\left[1 - \left(1 - \frac{1200}{30000}\right)^{300}\right]^{150} > 0.99928,$$

which is conservative since the inner exponent is really varying from 449 to 300.

Randomly selecting such a *walk* across the clone, the probability that the walk reaches the other end of the clone is greater than

$$2\binom{150}{50}\left(\frac{1}{2}\right)^{150} > 3 \times 10^{-5}.$$

This conservatively estimates the probability that at least 50 of the successive overlaps begin in the right-justified half of the 1200 length region of the previous fragment (and so extend the walk by at least 600 bases). Thus the probability that the selected walk covers the clone is greater than 0.999.

Sequencing the fragments from both ends yields the sequence, assuming read lengths of 650. The advantage of longer reads is that longer fragments can be used and hence for a desired probability for coverage, fewer fragments can be used. A distinct possibility is that merely improving the *percentage* of long reads has a significant effect.

We emphasize that these are simply lower bounds which are rather conservative, computed for this idealized case.

We next consider the probability that a complete tiling can be constructed and correctly assembled given a specific error rate in the base calls. Suppose that G is a sequence of bases of length x, G^* is a probabilistic garbling of G with an error rate $1-E$ and that R is a random sequence of length x. For each $m < x$, the probability that G^* and G disagree in at most m places is

$$p_m = \sum_{k=0}^{m} \binom{x}{k} E^k (1-E)^{x-k},$$

while the probability that R and G disagree in at most m places is

$$q_m = \sum_{k=0}^{m} \binom{x}{k} (0.75)^k (0.25)^{x-k},$$

which is dominated by the last term for the relevant values of x and m.

We examine the case when x=100 and E=0.1. In the assembly problem, p_m should be calculated with a smaller error rate since one is considering matches between two garbled sequences. For an error rate of $E = 0.1$, the effective error rate is approximately 0.186. Typical values for varying choices of m are

$$p_{39} = 0.9999996; p_{40} = 0.99999987; p_{41} = 0.99999995.$$

The corresponding values for q_m are

$$q_{39} = 2.87 \times 10^{-14}; q_{40} = 1.33 \times 10^{-13}; q_{41} = 5.90 \times 10^{-13}.$$

At each stage of the construction of the walk and with a threshold of m, the probability that there is an assembly error which passes the threshold requirement is at most

$$1 - (1 - q_m)^{1200\times 450}.$$

The probability that a correct fragment will pass, correctly placed, is at least p_m (in the worst case of there only being one such fragment). Thus, if there is a walk across the clone, the probability of constructing a valid walk across the clone is at least

$$P_m = (1 - q_m)^{1200\times 450\times 150} \times {p_m}^{150}.$$

With values as above, we have

$$P_{39} = 0.99993; P_{40} = 0.99997; P_{41} = 0.99996.$$

With a threshold of 40 the probability of constructing a correct walk across the clone is essentially the same (0.999) as the probability that there exists such a walk across the clone.

The analysis here makes several (important) simplifying assumptions. For example, it assumes that the fragments are uniformly distributed across the clone and that the clone itself is a random sequence of base pairs. While in some regions of the genome the latter may be a good assumption, there are certainly areas where it is not. More importantly, even somewhat limited partial repeats within the clone will have a possibly significant impact on the analysis. This can be explored experimentally via computer simulations using known stretches of the sequence (Section 3.3.1).

Further, with fragments produced using sets of restriction enzymes, the fragments may well not be uniformly distributed and we only considered pointwise garbling (not insertions or deletions). However the intent of this analysis is simply to illustrate the relative importance of base-calling accuracy and coverage (number of fragments) in the sequencing process.

Another important point is that attention should be paid to examining the relative merits of:

- Having the sequence of the genome at relatively low accuracy, together with a library of fragments mapped to the sequence;
- Having the sequence of the genome at high accuracy.

There are sequencing strategies in which the order of the fragments is essentially known in advance. The assembly of such a library of fragments is easier (significantly easier for the idealized random genome). It is possible that for sequencing certain regions of the genome these approaches, coupled to accepting higher error rates in the reads, are superior.

A final point concerning accuracy is the placement of known sequences against the garbled genome sequence. Suppose that, as above, the garble rate is 0.1; i.e., the accuracy is 0.9. Then given a sequence of length 50 from the true sequence, the probability that the sequence is correctly, and uniquely, placed is 0.999 using a threshold of 12 errors. Again, the assumptions are that the genome sequence is random or at least that the given segment is from a portion of the genome which is random. However if a significant fraction of the genome is random then (with high probability) false placements will only happen in the remaining fraction of the genome. This could be used to produce interesting kinds of maps, using a small library of target fragments. Again, some simulations can easily test these various points against known sequence data and allowing errors of insertion and deletion.

3.2 Verification Protocols

Since the "proof of the pudding" lies in the actual accuracy of the output, absolute accuracy can be determined only by physical testing of the sequence output. That is, given the putative sequence of base pairs for a certain contig (which we term the "software sequence"), independent protocols should be established to verify this software sequence relative to the physical contig. Such "verification" is a different task from *de novo* sequencing itself, and should be accomplished by means as independent as possible from those employed to obtain the initial sequence.

An ideal verification method would be:

- **Sequence blind**: requires no *a priori* knowledge of the sequence
- **Sequence independent**: efficacy independent of the sequence being verified
- **Reliable**: a high probability of detecting errors, with low probability of false alarms
- **Economical**: cost (labor, materials, time) a small fraction of the cost of sequencing
- **Capable**: long sequences easily verified
- **Specific**: provides further information about the errors beyond "Right or Wrong"

One obvious strategy is to resequence the DNA by a method different than that used by the original researcher. Unfortunately, this fails on the grounds of economy and the fact that today there is really only one large-scale sequencing technique.

In this section, we describe two possible verification protocols, and close with a discussion of the implementation of *any* protocol.

3.2.1 Restriction enzyme verification of sequence accuracy

We propose Multiple Complete Digestions (MCD) as a verification protocol satisfying most of the criteria above. It will allow statements like "With

90% probability, this sequence is accurate at the 10^{-3} level" or, more generally, "With confidence C, the sequence is accurate at the $\mathcal{E}$ level." It may also be used to localize and characterize errors in the sequence.

MCD has been developed and used as a method for generating high-quality physical maps preparatory to sequencing [G. K. S. Wong *et al.*, PNAS **94**, 5225–5230, 1997]. Here, we quantify the ability of this technique to provide probabilistic sequence verification.

The basic idea is that the putative sequence unambiguously predicts the fragment lengths resulting from digestion by any particular endonuclease, so that verification of the fragment lengths is a necessary (but not sufficient) check on the sequence. Multiple independent digestions then provide progressively more stringent tests. Of course, if the putative sequence has been generated by MCD with one set of enzymes, a completely different set must be used for verification.

Let us assume that $\mathcal{E}$ is the single-base error rate, that only single-base substitutions or deletions can occur, and that we are using restriction enzymes specific to a b-base pattern (most commonly, $b = 6$ for the enzymes used in sequencing, although enzymes with $b = 4, 5, 7$, and 8 are also known).

A digestion will give an error (i.e., fragments of unexpected length) when an error has destroyed a restriction site or created a new one from a "near-site" of b-bases whose sequence differs from the target sequence by one base (we ignore the probability of two or more errors occurring simultaneously within a restriction site or near-site). Then the probability of any one restriction site being destroyed is $b\mathcal{E}$ (since the error can occur in any one of the b positions), while the probability of a near-site being converted is $\mathcal{E}/3$ (since only one of the three error possibilities for the "wrong base" leads to a true site).

Then the expected number of errors in a sequence containing S sites and N near sites is

$$\langle E \rangle = \mathcal{E}bS + \mathcal{E}N/3 \equiv \mathcal{E}\sigma$$

where $\sigma = bS + N/3$ is the effective number of sites.

3.2.1.1 *Probabilistic estimate*

Let us now consider a sequence of length L bases. Assuming that bases occur at random, we expect $S = L/4^b$ sites for a single restriction enzyme and $N = 3bL/4^b$ near sites, since there are 3 ways each of the b bases at a site can differ from the target pattern. Hence, for D different digestions, we expect

$$\sigma = 2DbL/4^b.$$

Since the number of fragments expected if there are no errors is $S = L/4^b$ and a convenient number of fragments to separate is $S = 10$, taking b=6 implies a sequence length of $L = 40$ kb (the size of cosmid clones) and $\sigma = 120D = 600$ if $D = 5$.

3.2.1.2 *Real DNA*

The probabilistic estimate of σ above assumed that all b-mers were equally likely, or more precisely, that the recognized b-mers were uniformly distributed. However, there is no need to make that assumption when DNA is presented for checking. Instead one can scan the proposed sequence and count the number of sites where errors could make a difference in how the sequence is cleaved. The calculation mimics exactly the random model above: each recognized site contributes 1 to σ and each near site contributes 1/3. The total for the sequence is then the contribution of that endonuclease to σ.

The table below shows the results of this counting for $D = 5$ restriction enzymes for three pieces of human sequence from the Whitehead Center: L10 of length 48 kb, L8 of length 47 kb, and L43 of length 44 kb. (The first two are on 9q34, while the third is on the Y chromosome). Also considered is a completely random sequence of 40 kb.

Site Fragment	L10 (48 kb)	L8(47 kb)	L43(44 kb)	Random (40 kb)
GGATCC *(BamI)*	126	117	112	137
GATATC (*Eco*RV)	49	40	105	94
AAGCTT *(HindIII)*	66	112	134	121
TCTAGA (*BglII*)	84	79	190	145
TGGCCA *(MscI)*	295	377	109	122
σ	620	725	650	619

These results agree with the probabilistic estimate of $\sigma \sim 600$ for $D = 5$ and $L \sim 40$ kb. However, while the probabilistic model is true on average, it is not true in detail and some restriction enzymes give more meaningful tests of a given sequence than others (i.e., contribute more to σ). For example, digestion of L10 with *Eco*RV does not add very much information, while digestion with *Msc*I does. Hence, for a given DNA sequence, it is possible to choose the most meaningful set of restriction enzymes to be used in the test.

3.2.1.3 *Judging the results*

When a particular sequence is digested with a particular set of enzymes, the number of errors actually observed will be given by a Poisson distribution, in which the probability of observing E errors is

$$P(E) = \frac{\langle E \rangle^E}{E!} e^{-\langle E \rangle}.$$

What can be learned from a MCD test that shows E errors? Let us assume that the tests are arranged so that σ=700, that $\mathcal{E}$=10^{-3} the quality goal, and that we declare that any sequence showing $E < 2$ errors in an MCD test is "good." In that case, there is a false alarm probability of P_{FA}=0.16 that an $\mathcal{E}$=.001 sequence will be rejected, and will have to be redone. However, if the sequence has $\mathcal{E}$=0.01, there is only a P_A=0.007 probability that it will be accepted. Hence, this simple operational definition (at most one error) implies only slightly more work in resequencing, but gives high confidence ($> 99\%$) in a sequence accuracy at the level of $\epsilon = 0.01$ and 90% confidence in the sequence at the $\mathcal{E} \sim 0.005$ level. The implications of other choices for the maximum acceptable number of errors or for different values of $< E >$ follow straightforwardly from the properties of the Poisson distribution; some representative values for σ=700 are given in the table below.

	E<1	E<2	E<3	E<4
$P_{FA}(\mathcal{E}$=0.001)	0.50	0.16	0.035	0.006
$P_A(\mathcal{E}$=0.01)	0.0009	0.007	0.03	0.08
$\mathcal{E}$ ($P_A = 0.1$)	0.003	0.005	0.008	0.010

Note that the estimates above assume both perfect enzyme specificity; and sufficient fragment length resolution (1% seems to be achievable in practice, but one can imagine site or near-site configurations where this would not be good enough, so that a different set of restriction enzymes might have to be used). The extent to which these assumptions hinder MCD verification, as well as the ability of the method to constraint sequence to $\mathcal{E}< 10^{-4}$, can best be investigated by trials in the laboratory.

3.2.2 Hybridization arrays for sequence verification

As we have discussed in Section 2.3.3, the combinatorics make *de novo* sequencing a formidable challenge for present-day hybridization arrays. However, beyond the differential sequencing applications we have discussed, one potentially important application of hybridization arrays is to the problem of sequence quality control and verification, particularly since it is extremely important to employ means independent of those used to derive the putative sequence of a particular contig.

Hybridization arrays could provide a method for sequence verification independent of the present Sanger sequencing. The strategy would be to construct a Format 2 array based upon the candidate sequence for the contig. This array would then be challenged by the physical contig, with the goal being to detect differences between the "software" sequence as determined by a previous sequencing effort and the "hardware" sequence of the contig itself. For this protocol the "software" sequence would be represented by the oligomer probes of the array. Since the objective is to detect differences between two very similar sequences, the requirements on the number of distinct probes and hence on the size of the array are greatly relaxed as compared to the previous discussion of *de novo* sequencing. More explicitly, to scan a target contig of length N bases for single-base mismatches relative to a "known" (candidate) sequence, an array of $4N$ probes is required, which would increase to $5N$ if single site deletions were included. The array might include as well sets of probes designed to interrogate specific "problem" sections of the target. For $N \sim 40$ kb, the required number of probes is then of order 2×10^5, which is within the domain of current commercially capability.

Note that relative to the proposal in Section 3.3.2 to establish "gold standards" of DNA sequence, this strategy could also play an important role in helping to verify independently the standards themselves.

A case study relevant to the objective of sequence verification and error detection by hybridization is the work of M. Chee *et al.* [op cit.], for which an array with 135,000 probes was designed based upon the complete (known) 16.6 kb sequence of human mitochondrial DNA. As illustrated in Figure 3.1, this work detected sequence polymorphisms with single-base resolution, with 15-mer probes. Note that the total number of probes (135,000) is considerably smaller than the total possible set for a 15-mer ($4^{15} \sim 10^9$), allowing considerable flexibility in the design of the probes. In terms of an overall figure of merit for accuracy, the simplest possible procedure was employed whereby a scan to detect the highest fluorescent intensity from among

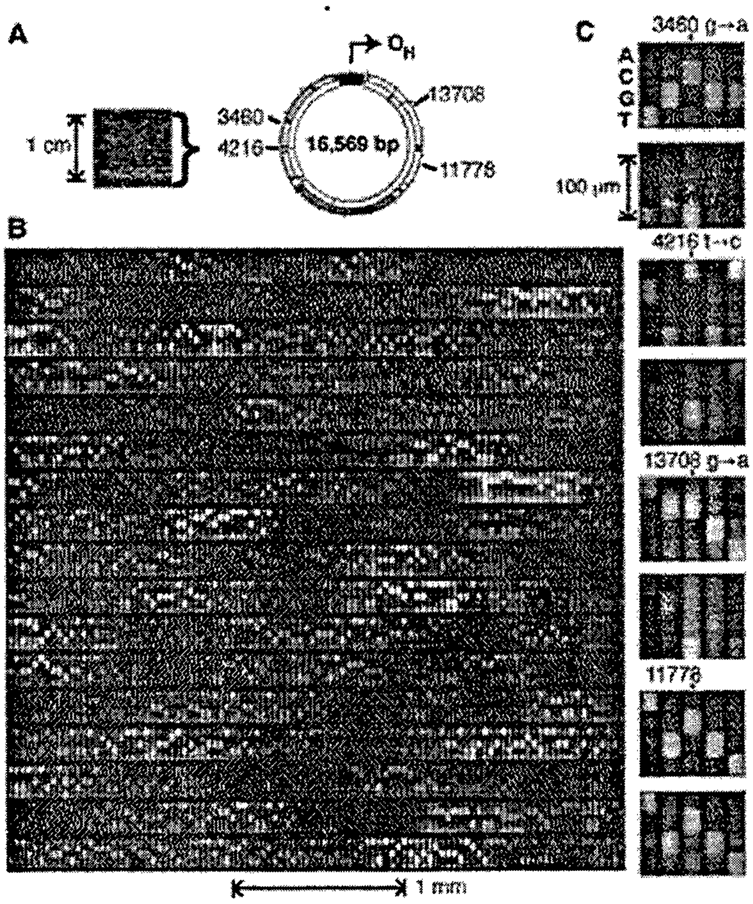

Figure 3.1. Human mitochondrial genome on a chip. (A) An image of the hybridized array with 135,000 probes designed to interrogate the 16.6 kb mitochondrial target RNA. (B) A magnified portion of the array. (C) Illustration of the ability to detect single base-pair differences. [from M. Chee *et al.*, op cit.]

the four possible base substitutions was made and led to 99% of the target sequence being read correctly. While this accuracy is not overwhelming, considerable improvement could presumably be achieved by incorporating more sophisticated analysis algorithms which take into account the overall pattern of mismatches, such as the were in fact employed by Chee et al. in their studies of polymorphisms for mitochondrial DNA from various populations. Of course since DNA is eubacterial in character, many of the more challenging sequence pathologies are absent relative to eukaryotic DNA. Still, Chee *et al.* provides a useful benchmark against which to assess the potential of hybridization arrays for sequence verification.

Perhaps the most important motivation for suggesting this strategy for verification is that the "mistakes" associated with sequence determination from target-probe interactions in a massively parallel fashion may well be sufficiently different from those arising from the gel-based procedures so as to give an independent standard for accuracy. Of course there are a host of issues to be explored related to the particular kinds of errors made by hybridization arrays (including the fidelity with which the original array is produced, hybridization equivalents, etc.). For the purpose at hand, attention should be focused on those components that most directly impact the accuracy of the comparison.

Particular suggestions in this regard relate to the readout and image processing for the array, tasks which are often accomplished site by site via scanning confocal microscopy. It would seem that alternate readout strategies should be explored, including (perhaps image-intensified) CCDs. Since the target sequence is known with small uncertainty as are the set of errors associated with single-base substitutions and deletions as well as with other "typical" errors in sequence reconstruction, image templates could be precomputed and cross-correlated with the actual image by adapting algorithms from the image processing community to improve the accuracy with which information is extracted from the array.

The value of such a strategy for sequence verification extends beyond that of providing an independent avenue for error checking. It would also couple the traditional gel-based effort to emerging advanced technologies, with potential benefit to both. Moreover, it could be part of a broader attempt to define a longer-term future for the sequencing centers as new technologies come on line to supplant gel-based sequencing and as emphasis shifts from *de novo* sequencing to comparative studies such as related to polymorphisms.

3.2.3 Implementation of verification protocols

Any verification protocol must require significantly less effort than resequencing, and so there will be considerable latitude in its implementation. In one limit, sequencing groups might be required to perform and document verification protocols for all finished sequence that they wish to deposit in a database. Alternatively, a "verification group" could be established to perform "spot" verifications of database entries selected at random. A third possibility is to offer a "bounty" for identifying errors in a database entry.

Clearly, economic, sociological, and organizational factors must be considered in choosing among these, and other, possible implementations. We recommend that the funding agencies promote a dialog within the sequencing communities about possible verification protocols and their implementation.

3.3 Assessing and Improving Present Techniques

Our emphasis on quantitative metrics for accuracy is made against the backdrop of inadequate understanding of the quality of the "end product" in the current Human Genome sequencing effort. While the level of competence and effort devoted to "doing the job right" in the sequencing centers is commendable, there is a clear need to implement a comprehensive program of quality assurance and quality control. Here we suggest some ways to provide more quantitative measures of the errors in the end product, and to understand how the various steps in sequencing contribute to the overall error budget.

Quality assurance and quality control (QA/QC) are of sufficient importance to be made integral requirements in the Project. Each sequencing center should invest a fraction of its own budget to characterize and understand its particular accuracy and error rates. This should be part of a continuing effort, rather than a one-time event. Quality control within the Centers should not be externally micro-managed, but each Center should be required to develop its own credible plan for QA/QC.

We further urge that the effort to develop new QA/QC technology be tightly coupled to the sequencing centers. In particular, new technologies such as large scale hybridization arrays or single-molecule sequencing are not currently competitive with gel-based electrophoresis for high throughput sequencing and long base reads, but they could be exploited in the short term for "niche" applications such as sequence verification for QA/QC. In the longer term, the Centers must integrate new technical advances into their operations, and the avenue of QA/QC is an important mechanism to help to accomplish this goal. From a longer-term perspective it seems clear that after the human genome has been sequenced once, emphasis will shift toward differential sequencing relevant to the study of polymorphism and homologies, and to the genetic origins of disease. QA/QC can thus be viewed as part of a broader effort to define a long-term future for the sequencing Centers, with technological leadership at the forefront as a prime component.

3.3.1 A systems approach is required

As we have noted, the issues of accuracy and error rates in reconstructed genomic information are crucial to the value of the "end-product" of the Human Genome Project, yet requirements for accuracy are complex and detail-dependent. DOE should support a portfolio of research in genome quality assurance and quality control issues.

One of the elements of this research should be computer simulation of the process of sequencing, assembly, and finishing. We believe that research into the origin and propagation of errors, *through the entire system*, are fully warranted. We see two useful outputs from such studies: 1) more reliable descriptions of expected error rates in final sequence data, as a companion to database entries, and 2) "error budgets" to be assigned to different segments of mapping and sequencing processes to aid in developing the most cost-effective strategies for sequencing and other needs.

DOE should solicit and support detailed Monte Carlo computer simulation of the complete mapping and sequencing processes. The basic computing methods are straight-forward: an reference segment of DNA (with all of the peculiarities of human sequence, such as repeats) is generated and subjected to models of all steps in the sequencing process; individual bases are randomly altered according to models of errors introduced at the various stages; the final, reconstructed segment or simulated database entry is compared with the input segment and errors are noted.

Results from simulations are only as good as the models used for introducing and propagating errors. For this reason, the computer models must be developed in close association with technical experts in all phases of the process being studied so that they best reflect the real world. This exercise will stimulate new experiments aimed at the validation of the error-process models, and thus will lead to increased experimental understanding of process errors as well.

Useful products of these types of simulations are "error budgets" for different steps in the measurement and analysis chain. Such budgets reflect the contributions of individual steps and their effect on the accuracy of the final result. This information can be used, for example, to establish quality criteria for the various stages of the sequencing process, so that emphasis and funds can be devoted to improving the accuracy of those steps which have the strongest influence on the accuracy of the final sequence product.

Error budgets will depend on the final accuracy required for a specific use of the end-product, which is analyzed sequence information. By comparing specific end-product needs for accuracy and quantity of information with error budgets and costs of individual steps in the overall process from DNA to database, it should be possible to perform cost/benefit analyses for developing optimum sequencing strategies.

3.3.2 Gold standards for measuring sequence accuracy

DOE should take the lead in developing "gold standards" for human DNA sequence. Standard DNA sequences could be used by the whole sequencing community for assessing the quality of the sequence output and sequencing protocol through "blind" experiments within the various centers. These gold standards should be designed to highlight quality assessment in "hard" DNA-sequencing regions and in potential problem areas, as well as in "ordinary" coding regions. They would consist of cloned DNA molecules of two types:

- a cosmid vector containing an insert of ~40 kb of human DNA that has been sequenced with high accuracy and assembled without any unresolved discrepancies;
- a phagemid vector containing an insert of ~1 kb of synthetic DNA including both human-derived sequences and contrived sequences that are known to cause common artifacts in DNA sequence acquisition.

The standard cosmid will have been transduced and propagated in bacterial cells, then stored as individual aliquots kept at -70°C. Upon request, one or more of these aliquots would be made available to a sequencing group. All of the subsequent steps, including further propagation of the cosmid, restriction mapping, subcloning, sequencing, assembly, and finishing would be carried out by the sequencing group. Performance could be assessed based on a variety of indices such as PHRED and PHRAP scores, number of sequencing errors relative to the known standard, type of sequencing errors, time required to complete the sequencing, and cost of sequencing. The cosmid standard might also be used to compare alternative sequencing protocols within a sequencing center or to conduct pilot studies involving new instrumentation.

The standard phagemid will have been produced in large quantity, purified, and stored as individual aliquots kept at -70°C. After thawing, the DNA will be ready for sequencing, employing "universal" primers that either accompany the template DNA or are provided by the sequencing group. The purpose of this standard is to assess the quality of DNA sequencing itself, based on indices such as PHRED score, read length, and the number and type of sequencing errors relative to the known standard. The target sequence will have been designed to elicit common sequencing artifacts, such as weak bands, strong bands, band compressions, and polymerase pauses.

Although the cosmid standard is expected to have greater utility, the phagemid standard will be used to control for variables pertaining to DNA sequencing itself within the overall work-up of the cosmid DNA. It is likely that the sequencing groups will be on their "best behavior" when processing a gold standard, resulting in enhanced performance compared to what might be typical. This cannot be avoided without resorting to cumbersome procedures such as surprise examinations or blinded samples. Thus it will be important to examine not only the output of the sequencing procedures, but also the process by which the data is obtained. The extent to which it is possible to operate in a "best behavior" mode will itself be instructive in assessing DNA sequencing performance. At the very least, such trials will establish a lower limit to the error rate expected.

We recommend that the DOE provide funding, on a competitive basis, to one or two individual investigators who will construct and maintain the DNA standards. It might be appropriate to construct a small family of cosmid and phagemid standards that would be made available sequentially. The experience of the sequencing groups in processing these gold standards will suggest ways in which they could be improved to better assess critical aspects of the sequencing process.

3.3.3 Quality issues pertaining to sequencing templates

While most of our discussion has involved QA/QC issues in the sequencing and assembly process, it is useful to consider also quality issues in the processes used to prepare DNA for sequencing. We do so in this subsection.

There are many steps involved in construction of a human genomic DNA library and subcloning of that library into a form suitable for automated DNA sequencing. These include:

1. fragmentation of chromosomal DNA by mechanical shearing or partial enzymatic digestion;
2. size fractionation of the DNA fragments by gel electrophoresis or centrifugation;
3. cloning of ~1 Mb fragments into high-capacity vectors, such as YACs or BACs;
4. propagation of YACs or BACs within host cells;

5. enzymatic digestion of YAC or BAC inserts to obtain fragments of ~40 kb;
6. cloning into medium-capacity cosmid vectors;
7. propagation of cosmids within bacterial cells;
8. enzymatic digestion of cosmid inserts to obtain fragments of ~1 kb;
9. cloning into low-capacity plasmid or phagemid vectors;
10. preparation of purified plasmid or phagemid DNA.

The subsequent steps of dideoxy sequencing, base calling, assembly, and finishing are all prone to errors that can be investigated and quantified, as we have discussed in previous sections. However, each of the steps above can introduce artifacts that make sequencing more difficult.

The steps involved in the preparation of templates for sequencing are made error-tolerant by the exponential amplification that is inherent in these procedures. Errors do occur, such as empty vectors, poor transformation efficiency, insufficient vector amplification, and inadequate purity of the template DNA. These problems usually result in clones that drop out of the process. Provided that there is redundant coverage of the DNA among the successful clones, the failed clones can essentially be ignored. However three quality control issues pertaining to template preparation merit special attention:

1. There may be incomplete representation of the genomic DNA at the level of the BAC/YAC, cosmid, or plasmid/phagemid libraries. This may be due to insufficient redundancy in construction of the library, but more often they are due to regions of the chromosome that are either difficult to clone or difficult to propagate within host cells. The genomics community is well aware of these problems and has taken appropriate countermeasures. Unlike the yeast genome, which has

been sequenced successfully in its entirety, there may be regions within the human genome that cannot be cloned and therefore cannot be sequenced. At present the best course of action is to press ahead and deal with the problem of "unsequenceable" DNA if and when it arises.

2. There may be spurious DNA sequences intermixed with the desired genomic DNA. The two most common sources of contamination are vector-derived DNA and host cell DNA. Vector sequence can be recognized easily by a suitable sequence-matching algorithm. Incredibly, there are many entries in the genomic databases today that are either partly or completely derived from vector sequence. Host cell DNA is more difficult to recognize, but these too can be identified with the complete genomic sequences of yeast and *Ecoli* available. Although spurious sequences can be eliminated after the fact, it should be made incumbent on the sequencing centers to do this prior to database submission.

There are challenges in maintaining proper inventory control over the vast number of clones and subclones that are being generated by the Project. Current procedures at the major genome centers are adequate in this regard. A physical inventory should be maintained for all BAC/YAC and cosmid clones, but this is not critical for the plasmid/phagemid clones. An electronic inventory, with secure back-up copies, should be maintained for all clones and subclones that are generated.

4 GENOME INFORMATICS

4.1 Introduction

In a statement of research goals of the US Human Genome Project [F. Collins and D. Galas, "A new five-year plan for the US Human Genome Project," Science 262: 43–46 (1993)], the Project's leaders define "informatics" as:

> *.....the creation, development, and operation of databases and other computing tools to collect, organize, and interpret data.*

Their goals for the current 5-year period are:

- Continue to create, develop, and operate databases and database tools for easy access to data, including effective tools and standards for data exchange and links among databases.
- Consolidate, distribute, and continue to develop effective software for large-scale genome projects.
- Continue to develop tools for comparing and interpreting genome information.

While similar in purpose and style to other major scientific cataloging efforts of the past and present—for example, Handbook of Chemistry and Physics, Chart of the Nuclides, to name two familiar resources—the Human Genome Projects informatics task is strikingly unified in that its focus is solely on translating and disseminating the information coded in human chromosomes. Genome informatics differs from earlier scientific catalogs also because it is a "child" of the information age, which brings clear advantages and new challenges, some of which are related to the following:

- the large amount of information to be assembled in meaningful ways, while the semantic content of that information is still largely not understood.
- the reliance on software algorithms at all stages from assembling pieces of information to interpreting results
- the large, globally distributed and diverse provider/user base
- the broad range of quality of information being processed and accessed, with uncertainties in even the measures of quality
- the rapidity with which the quantity and quality of information is increasing

Within the Human Genome Program, technical challenges in the informatics area span a broad range. Genome informatics can be divided into a few large categories: data acquisition and sequence assembly, database management, and genome analysis tools. Examples of software applications within the three categories include:

Data acquisition and sequence assembly:

- Process management and inventory control within Genome Centers
- Tools to track the pedigree of raw input data sources
- Servo control systems for individual robotic processes
- Software environments for coordinated distributed computing (e.g. robotic control systems) within a Genome Center
- Physical mapping software tools

- Base-calling software
- Sequence assembly tools
- Annotation tools; software for automatic sequence annotation
- Tools for automated submission of information to database centers

Database management:

- Local special-purpose databases
- Community-wide relational databases
- Software for database curation and quality control
- User "front ends" and interfaces for complex database queries
- "Middleware" for integration between separate databases
- Software to resolve semantic and nomenclature conflicts

Genome analysis:

- Data-mining tools
- Homology searches
- Identification of coding regions and genes
- Comparative genomics tools
- Placing proteins into gene families
- Tools for lineage analysis
- Tools for combinatorial analysis of hybridization array data

Managing such a diverse informatics effort is a considerable challenge for both DOE and NIH. The infrastructure supporting the above software tools ranges from small research groups (e.g. for local special-purpose databases) to large Genome Centers (e.g. for process management and robotic control systems) to community database centers (e.g. for GenBank and GDB). The resources which these different groups are able to put into software sophistication, ease of use, and quality control vary widely. In those informatics areas requiring new research (e.g. gene finding), "letting a thousand flowers bloom" is DOE's most appropriate approach. At the other end of the spectrum, DOE and NIH must face up to imposing community-wide standards for software consistency and quality in those informatics areas where a large user community will be accessing major genome data bases.

The need for genome quality assurance enters the informatics field at several different levels. At the earliest level, both policies and tracking software are needed that will preserve information about the pedigree (origin and processing history) of data input to the sequencing process. This potentially includes information on the origins of clones and libraries, image data of gel runs, and raw data of ABI-machine traces. Policies need to be developed concerning minimum standards for archiving the raw data itself, as well as for the index that will allow future users to find raw data corresponding to the heritage of a specific DNA sequence.

At the level of sequencing and assembly, DOE and NIH should decide upon standards for the inclusion of quality metrics along with every database entry submitted (for example PHRED and PHRAP quality metrics, or improvements thereon).

At the level of database quality control, software development is needed to enhance the ability of database centers to perform quality checks of submitted sequence data prior to its inclusion in the database. In addition, thought needs to be given towards instituting an ongoing software quality assurance program for the large community databases, with advice from appropriate commercial and academic experts on software engineering and quality control. It is appropriate for DOE to insist on a consistent level of documentation, both in the published literature and in user manuals, of the methods and structures used in the database centers which it supports.

At the level of genome analysis software, quality assurance issues are not yet well posed. Many of the current algorithms are highly experimental and will be improved significantly over the next five years. Tools for genome analysis will evolve rapidly. Premature imposition of software standards could have a stifling effect on the development and implementation of new ideas. For genome analysis software, a more measured approach would be to identify a few of the most promising emerging analysis tools, and to provide funding incentives to make the best of these tools into robust, well-documented, user-friendly packages that could then be widely distributed to the user community.

4.2 Databases

Currently, there are many, diverse resources for genomic information, essentially all of which are accessible from the World Wide Web. Generally, these include cross references to other principal databases, help-files, software resources, and educational materials. The overall impression one gets after a few hours of browsing through these web sites is that of witnessing an extraordinarily exciting and dynamic scientific quest being carried out in what is literally becoming a world-wide laboratory.

Web tools and the databases are also changing how the biology community conducts its business. For example, most journals now require a "receipt" from one of the standard databases indicating that reported sequence data have been filed before a paper is published. The databases are finding ways to hold new entries private pending review and publication. The databases contain explicit reference to contributors—there is probably no better way to exercise real quality control than the threat of exposure of incorrect results. We view all these developments as being very positive.

With so much information coming available, considerable effort goes into staying current. Many institutions conduct daily updates of information from the database centers. This works because such updates can be performed automatically off of peak working hours. The resources needed to update and circulate information are likely to increase as volume increases. The effort in learning how to use relevant database tools represents an important investment for individual scientists and group leaders.

Maintenance of databases is an important resource question for the Project. Currently, DOE supports two major efforts:

1. **Genome Sequence DataBase (GSDB)** (www.ncgr.org) operated by the National Center for Genome Resources which was established in Santa Fe in July, 1994. GSDB is described in its Web information as "one of the key components of the emerging federated information infrastructure for biology and biotechnology."

2. **The Genome Database (GDB)** (www.gdb.org) was established at Johns Hopkins University in Baltimore, Maryland in 1990. GDB is the official central repository for genomic mapping data resulting from the Human Genome Initiative. In support of this project, GDB stores and curates data generated worldwide by those researchers engaged in the mapping effort of the Human Genome Project.

GenBank (www.ncbi.nlm.nih.gov/Web/Genbank/index.html) is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 967,000,000 bases in 1,491,000 sequence records as of June 1997. GenBank is part of the International Nucleotide Sequence Database Collaboration, which also includes the DNA Data

Bank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL/EBI) Nucleotide Sequence Database.

4.2.1 User issues

The major genomic databases serve broad communities, whose users have vastly differing needs. In this situation several levels of user input and management review are called for.

To assure that all the database centers are "customer oriented" and that they are providing services that are genuinely useful to the genome community, each database center should be required to establish its own "Users Group" (as is done by facilities as diverse as NSFs Supercomputer Centers and NASA's Hubble Space Telescope). Membership in these "Users Groups" should be on a rotating basis, and should represent the full cross-section of database applications (small academic groups, large genome centers, pharmaceutical companies, independent academic laboratories, etc.). The "Users Groups" should be convened by each Center Director and should meet several times a year, with written reports going to the Center Directors as well as to the sponsoring Federal agencies.

Several briefers from database centers expressed concern that the "average user" was not well-informed about appropriate ways to query the databases, and that search tools (e.g. BLAST) frequently were not being used in a sound fashion. To address this type of issue, DOE should encourage the database centers in consultation with their "Users Groups" to organize appropriate tutorials and workshops, to develop "crib sheets" and other instructional documentation, and to take further steps to educate targeted user communities in techniques for sound database use appropriate to their applications.

At a higher management level, DOE and NIH should continue the process of constituting independent panels every few years, to review the health of the entire suite of genomic database centers. These panels should provide independent peer review of every community database, including input from "Users Groups" as well as technical and management review of center operations. Inclusion of Computer Science database experts on the review panels will help facilitate exchange of information with the Computer Science community.

4.2.2 Modularity and standards

Too often database efforts attempt to "do it all"; i.e., they attempt to archive the data, provide mechanisms for cataloging and locating data, and develop tools for data manipulation. It is rare that a single data base effort is outstanding in all three areas, and linking the data too closely to the access and analysis methods can lead to premature obsolescence. For reference, the following functions can be identified:

Authoring: A group produces some set of data, e.g. sequence or map data.

Publishing and archiving: The data developed by individual authors is "published"electronically (i.e. put into some standard format) and accumulated in a network accessible location. This also involves some amount of "curation", i.e. maintenance and editing of the data to preserve its accessibility and accuracy.

Cataloging (metadata) : This is the "librarian" function. The primary function of a library is not to store information but rather to enable the user to determine what data is available and where to find it. The librarian's primary function is to generate and provide "metadata" about what data sets exist and how they are accessed (the electronic analog of the card catalogue). Other critical functions include querying, cross-referencing, and indexing.

Data access and manipulation: This is the "user interface". Because the data volumes are typically large, computerized methods for data access and manipulation must be provided, including graphical user interfaces (GUIs).

The key point is that the various functions should be modularized, rather than tangled together in a single monolithic effort. The reason is obvious: computer technology, storage technology, data base technology, networks, and GUIs are evolving on a time scale much shorter than the projected lifetime of the data. Each technology evolves on its own time scale and schedule. Therefore, the functions must be modularized to allow separate upgrading. Modularization also allows multiple approaches, e.g. to user access: simple, intuitive GUIs for some users, powerful search and combinatoric engines for others.

Data format standards are a key to successful modularity. The community should invest in developing a common "language" which includes definition of certain basic data types (e.g., "classes" or "objects" in object-oriented terminology). Data format conventions should be defined for sequence data, map data, etc. Where multiple standards already exist, investment should be made in translators. Some standardization of methods to operate on data objects is also desirable, particularly for the most frequent operations and queries. However, the user should be able to develop powerful customized methods and manipulation techniques.

Currently, neither standards nor modularity are very much in evidence in the Project. The DOE could contribute significantly by encouraging standards. Database groups should be encouraged to concentrate on the "librarian" functions, and leave the publishing and archival functions to other groups. Development of user interfaces and manipulation tools may also be tackled by database efforts, but it is not obvious that the best librarians are also the best GUI developers.

As part of the librarian function, investment should be made in acquiring automatic engines that produce metadata and catalogues. With the explosive growth of web-accessible information, it is unlikely that human librarians will be able to keep pace with the ancillary information on the genome, e.g. publications and web-sites. The technology for such search engines is well-developed for the web and needs to be applied specifically to genomic information for specificity, completeness, and efficiency.

Indexing and cross-referencing are critical database functions. It is often the case that the indexes which encapsulate the relationships in and between data bases constitute a far larger data set than the original data. Significant computer resources should go into pre-computation of the indexes that support the most frequent queries.

Consideration should be given by the database efforts to development of shell programs for genome database queries and manipulation. A shell is a simple interactive command-line interface that allows the user to invoke a set of standard methods on defined objects, and lists of objects. In the numerical world, Mathematica, Maple, and IDL are examples of such approaches. The shell typically has a simple syntax with standard if-then constructs, etc.

4.2.3 Scaling and storage

As noted in Section 1.2.3, about 40 Mb of human sequence data in contigs longer than 10 kb exists in the genome databases today, using a storage capacity of 60 GB (NCGR). By the time the Human Genome Project is complete, these databases can be expected to hold at least 3 Gb of sequence, along with annotations, links, and other information. If today's ratio of 1.5 KB per sequence-base is maintained, 4.5 TB of storage will be required. At the very least, a comparable 100-fold increase in submission/transaction rates will occur, but we expect the transaction rates to grow even faster as genomic data are more complete and searches become more sophisticated. While these capacities and transaction rates are well within the bounds of current database technology, careful planning is required to ensure the databases are prepared for the coming deluge.

4.2.4 Archiving raw data

As the Project proceeds it is reasonable to expect improvements in the analysis of the raw data. Therefore *a posterior* processing could be quite valuable, provided that *the trace data are archived.*

One of the algorithms used currently has been developed by P. Green. His base calling algorithm, PHRED, takes as input the trace data produced by the ABI instrument (chromatogram files). Quality parameters are developed based on qualitative features of the trace. Currently 4 such (trace) parameters are used. These are converted to quality thresholds through calibration on known sequence data.

Experiments conducted by Green, involving 17259 reads in 18 cosmids yielded the following results, comparing the error rates of the actual ABI software calling package to those of PHRED.

Method	**Sub**	**Del**	**Ins**	**Total**
ABI	4.26%	0.38%	1.47%	6.10%
PHRED	2.55%	0.58%	0.47%	3.60%

Of course, the distribution of errors should also be compared, error clusters have potentially serious implications for the assembly problem, more so than well isolated errors. Another potentially important consideration is the location of errors within the read.

It is not unreasonable to expect that the actual conversions, used in the PHRED algorithm, might be improved as the library of known sequence increases. Further, more than one conversion table might be required, depending on the general region of the genome one is attempting to sequence.

C. Tibbetts of George Mason University has developed a based calling algorithm based upon a neural network architecture. He has also worked to maximize the quality of the base calls through an engineering analysis of, for example, the ABI PRISM^{TM} 377.

Whatever algorithms are used it is important that the called sequence of bases have *associated confidence values* together with an interpretation of what these values are supposed to mean. For example confidence values could be pairs of numbers, the first representing the confidence that the base call is correct and the second representing the confidence that the base called is the *next base*. One might also consider adding a third coordinate representing the confidence that the called base corresponds to one base as opposed to more than one. These values should continually be checked for *internal* consistency; *every* read should be compared to the assembled sequence. This comparison involves the alignment of the read against the assembled sequence minimizing an adjusted error score.

Finally, there are currently several degrees of freedom in sequencing. Two, that could yield different (and hopefully independent) processes are:

1. Using dye labeled primers versus dye labeled terminators;

2. Sequencing the complementary strand.

Correlated errors define an upper bound in the accuracy of base calling algorithms that *cannot* be surmounted by repeated sequencing using the same chemistry. Ideally the confidence values assigned to individual base calls would closely correspond to these intrinsic errors. This can (and should) be tested experimentally.

There are two final points on the issue of archiving the raw data. More powerful algorithms (enhanced by either a growing body of knowledge about the genome or by better platforms) could improve the reads, and hence enhance overall accuracy. Such developments could also enable re-assembly in some regions (if they exist) where errors have occurred.

4.3 Measures of Success

Databases are crucial tools needed for progress in the Human Genome Project, but represent large direct costs in capital equipment and operations and potentially large hidden costs in duplication of effort and training. We believe the only true measure of success will be whether or not these tools are used by researchers making scientific discoveries of the first rank. That a given database installation is "better" than another in some theoretical sense is not sufficient. There are examples in consumer electronics where the "best" technology is not the one chosen by the majority—a similar situation could easily occur with databases in the Human Genome Project. We urge DOE to critically evaluate the "market impact" of the database efforts it supports by regularly surveying users and comparing with other efforts, supported outside DOE. Fundamentally, the operation of a major database is a service role—of very great importance and with real technical challenges—that may not be in the long-term interests of DOE, assuming other satisfactory database tools are available to its researchers at reasonable cost.

4.4 Sociological Issues

Until recently the biological sciences have been based upon relatively free-standing bench-top experimental stations, each with its own desk-top computer and local database. However a "sequencing factory" with high throughput faces new informatics needs: inventory management, a coordinated distributed computing environment (e.g. EPICS), automated tools for sequence annotation and database submission, and tools for sequence analysis. In addition the national and international Human Genome Projects must integrate the genomic information into a common and accessible data structure.

The broadly distributed nature of the Project presents a challenge for management of the informatics effort. In particular, across-the-board imposition of standards for software engineering and data quality will be difficult. The best course is for DOE to "choose its battles", emphasizing the development of common standards in areas of highest priority such as database centers, while tolerating a diversity of approaches in areas such as advanced algorithm development for genomic analysis. In addition to standards, consistent "User Group" input and peer review are needed for all of the genome database centers.

It will be helpful to increase the level of participation of the Computer Science community in genome-related informatics activities. While the human genome sequence database is not among the largest databases being developed today, the diverse nature of genome applications and the need to combine information from several different database sources provide real Computer Science challenges. Reaching out to the academic Computer Science community to engage the interest of graduate students and faculty members has not been easy to date. The genome community continues to debate whether it might be more fruitful to educate biologists in computer sciences, rather than educating computer scientists in biology. In our view, both approaches should continue to be pursued. DOE's informatics program should include outreach activities such as workshops, short courses, and other support which will familiarize Computer Scientists with the challenges of the genome program, *and* which will educate young biologists in those areas of Computer Science which are of importance to the genome effort.

DISTRIBUTION LIST

Director of Space and SDI Programs
SAF/AQSC
1060 Air Force Pentagon
Washington, DC 20330-1060

CMDR & Program Executive Officer
U S Army/CSSD-ZA
Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0150

Superintendent
Code 1424
Attn Documents Librarian
Naval Postgraduate School
Monterey, CA93943

Director
Technology Directorate
Office of Naval Research
Room 407
800 N. Quincy Street
Arlington, VA 20305-1000

DTIC [2]
8725 John Jay Kingman Road
Suite 0944
Fort Belvoir,VA 22060-6218

Dr A. Michael Andrews
Director of Technology
SARD-TT
Room 3E480
Research Development Acquisition
Washington, DC 20301-0103

Dr Albert Brandenstein
Chief Scientist
Office of Nat'l Drug Control Policy
Executive Office of the President
Washington, DC 20500

Dr H Lee Buchanan, I I I
Director
DARPA/DSO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Dr Collier
Chief Scientist
U S Army Strategic Defense Command
PO Box 15280
Arlington, VA 22215-0280

DARPA Library
3701 North Fairfax Drive
Arlington, VA 22209-2308

Dr Victor Demarines, Jr.
President and Chief Exec Officer
The MITRE Corporation
A210
202 Burlington Road
Bedford,M A01730-1420

Mr Dan Flynn [5]
OSWR
Washington, DC 20505

Dr Paris Genalis
Deputy Director
OUSD(A&T)/S&TS/NW
The Pentagon, Room 3D1048
Washington, DC 20301

Dr Lawrence K. Gershwin
NIC/NIO/S&T
7E47, OHB
Washington, DC 20505

Mr David Havlik
Manager
Weapons Program Coordination Office
MS 9006
Sandia National Laboratories
PO Box 969
Livermore, CA94551-0969

DISTRIBUTION LIST

Dr. Helmut Hellwig
Deputy Asst Secretary
(Science, Technology and Engineering)
SAF/AQR
1919 S. Eads Street
Arlington, VA 22202-3053

Dr Robert G Henderson
Director
JASON Program Office
The MITRE Corporation
1820 Dolley Madison Blvd
Mailstop W553
McLean, VA 22102

JASON Library [5]
The MITRE Corporation
Mail Stop W002
1820 Dolley Madison Blvd
McLean, VA 22102

Dr Anita Jones
Department of Defense
DOD, DDR&E
The Pentagon, Room 3E1014
Washington, DC 20301

Mr. O' Dean P. Judd
Los Alamos National Laboratory
Mailstop F650
Los Alamos, NM 87545

Dr Bobby R Junker
Office of Naval Research
Code 111
800 North Quincy Street
Arlington, VA 22217

Dr. Martha Krebs
Director
Energy Research
1000 Independence Ave, SW
Washington, DC 20858

Dr Ken Kress
Office of Research and Development
809 Ames Building
Washington, DC 20505

Lt Gen, Howard W. Leaf, (Retired)
Director, Test and Evaluation
HQ USAF/TE
1650 Air Force Pentagon
Washington, DC 20330-1650

Mr. Larry Lynn
Director
DARPA/DIRO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Dr. John Lyons
Director of Corporate Laboratory
US Army Laboratory Command
2800 Powder Mill Road
Adelphi,MD 20783-1145

Col Ed Mahen
DARPA/DIRO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Dr. Arthur Manfredi
ZETA Associates
10300 Eaton Drive
Suite 500
FairfaxVA 22030-2239

Dr George Mayer
Scientific Director
Army Research Office
4015 Wilson Blvd
Tower 3
Arlington, VA 22203-2529

Dr Bill Murphy
ORD
Washington, DC 20505

DISTRIBUTION LIST

Dr Julian C Nall
Institute for Defense Analyses
1801 North Beauregard Street
Alexandria, VA 22311

Dr Ari Patrinos [5]
Associate Director for
Biological and Environmental Reserach
ER-70
US Department of Energy
1901 Germantown Road
Germantown,MD 207874-1290

Dr Bruce Pierce
USD(A)D S
The Pentagon, Room 3D136
Washington, DC 20301-3090

Mr John Rausch [2]
Division Head 06 Department
NAVOPINTCEN
4301 Suitland Road
Washington, DC 20390

Records Resource
The MITRE Corporation
Mailstop W115
1820 Dolley Madison Blvd
McLean,VA 22102

Dr Victor H Reis [5]
US Department of Energy
DP-1, Room 4A019
1000 Independence Ave, SW
Washington, DC 20585

Dr Fred E Saalfeld
Director
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217-5000

Dr Dan Schuresko
O/DDS&T
Washington, DC 20505

Dr John Schuster
Technical Director of Submarine
and SSBN Security Program
Department of the Navy OP-02T
The Pentagon Room 4D534
Washington, DC 20350-2000

Dr Michael A Stroscio
US Army Research Office
P. O. Box 12211
Research Triangle NC27709-2211

Ambassador James Sweeney
Chief Science Advisor
USACDA
320 21st Street NW
Washington, DC 20451

Dr George W Ullrich [3]
Deputy Director
Defense Nuclear Agency
6801 Telegraph Road
Alexandria, VA 22310

Dr. David Whelan
Director
DARPA/TTO
3701 North Fairfax Drive
Arlington, VA 22203-1714

Dr Edward C Whitman
Dep Assistant Secretary of the Navy
C3I Electronic Warfare & Space
Department of the Navy
The Pentagon 4D745
Washington, DC 20350-5000